智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

面向道路交通场景的高效3D目标检测

陆军, 鲁林超, 翟晓阳, 刘霜

引用本文:
陆军, 鲁林超, 翟晓阳, 等. 面向道路交通场景的高效3D目标检测[J]. 智能系统学报, 2025, 20(1): 91-100.
LU Jun, LU Linchao, ZHAI Xiaoyang, et al. High-efficiency 3D object detection for road traffic scenes[J]. *CAAI Transactions on Intelligent Systems*, 2025, 20(1): 91-100.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202311013

DOI: 10.11992/tis.202311013
网络出版地址: https://link.cnki.net/urlid/23.1538.TP.20240921.1716.002

面向道路交通场景的高效 3D 目标检测

陆军，鲁林超，翟晓阳，刘霜
（哈尔滨工程大学 智能科学与工程学院，黑龙江 哈尔滨 150001）

摘　要：针对当前两阶段的点云目标检测算法 PointRCNN: 3D object proposal generation and detection from point cloud 在点云降采样阶段时间开销大以及低效性的问题，本研究基于 PointRCNN 网络提出 RandLA-RCNN(random sampling and an effectivelocal feature aggregator with region-based convolu-tional neural networks) 架构。首先，利用随机采样方法在处理庞大点云数据时的高效性，对大场景点云数据进行下采样；然后，通过对输入点云的每个近邻点的空间位置编码，有效提高从每个点的邻域提取局部特征的能力，并利用基于注意力机制的池化规则聚合局部特征向量，获取全局特征；最后使用由多个局部空间编码单元和注意力池化单元叠加形成的扩展残差模块，来进一步增强每个点的全局特征，避免关键点信息丢失。实验结果表明，该检测算法在保留 PointRCNN 网络对 3D 目标的检测优势的同时，相比 PointRCNN 检测速度提升近两倍，达到 16 f/s 的推理速度。

关键词：深度学习；3D 目标检测；点云；随机采样；局部特征聚合；注意力机制；自动驾驶
中图分类号：TP391　　**文献标志码**：A　　**文章编号**：1673−4785(2025)01−0091−10

中文引用格式：陆军，鲁林超，翟晓阳，等. 面向道路交通场景的高效 3D 目标检测 [J]. 智能系统学报，2025，20(1)：91–100.
英文引用格式：LU Jun, LU Linchao, ZHAI Xiaoyang, et al. High-efficiency 3D object detection for road traffic scenes[J]. CAAI transactions on intelligent systems, 2025, 20(1): 91–100.

High-efficiency 3D object detection for road traffic scenes

LU Jun, LU Linchao, ZHAI Xiaoyang, LIU Shuang
(College of Intelligent Systems Science and Engineering, Harbin Engineering University, Harbin 150001, China)

Abstract: Based on the 3D object proposal generation and detection from pointcloud, namely PointRCNN network, this study proposes an RandLA-RCNN architecture to address the issues of high time cost and inefficiency in the point cloud downsampling stage of the current two-stage point cloud object detection algorithm. Firstly, by taking advantage of the efficiency of random sampling method, the large-scale point cloud data are downsampled to handle massive point cloud data. Then, the spatial positions of each neighboring point of the input point cloud are encoded to effectively enhance the ability of each point to extract local features from the neighborhood. Attention-based pooling rules are used to aggregate local feature vectors and obtain global features. Finally, an extended residual module formed by stacking multiple local spatial encoding units and attention pooling units is used to further enhance the global features of each point and avoid the loss of key point information. Experimental results show that this detection algorithm retains the advantages of PointRCNN network in detecting 3D objects, while achieves nearly twice the detection speed compared with PointRCNN, reaching an inference speed of 16 frames per second.

Keywords: deep learning; 3D object detection; point cloud; random sampling; local feature aggregation; attention mechanism; autonomous driving

3D 传感器技术的迅速发展，使 3D 目标检测成为了当前计算机视觉领域主要的研究方向。当前国内外对于 3D 目标检测算法的研究根据输入模态的不同主要分为两个不同的方向，分别是基于多模态的目标检测算法的研究以及基于激光雷达点云的 3D 目标检测算法的研究[1]。

基于多模态输入的 3D 目标检测方法的共同特点是都会特别受图像数据和点云数据的融合方法的影响[2]。根据特征融合方式是否投影可以划

收稿日期：2023−11−13.　**网络出版日期**：2024−09−23.
基金项目：黑龙江省自然科学基金项目（F201123）.
通信作者：陆军. E-mail: lujun0260@sina.com.

分为基于投影的方法和基于非投影的方法[3]。基于投影的多模态 3D 目标检测方法在特征融合阶段使用投影矩阵来实现点云和图像特征的整合[4-6]，但是这种方法很容易受到投影矩阵质量的影响。而基于非投影的方法则不依赖特征对齐来实现融合[7-12]，其大都采用注意力机制或者构造统一特征空间的方法去对齐图像与点云的特征信息。虽然多模态的方法在一定程度上能够弥补单一传感器带来的缺陷，但是如何理解使用不同模态数据之间的关联与互补之处，是多模态方法在应用中的重要问题[13]。

由于点云不具有二维图像的规则性[14]，因此现有的二维目标检测方法无法直接使用。根据点云处理方式的不同，基于激光雷达点云的目标检测算法可分为 3 种类型：原始点云转换为体素格式后输入网络进行检测，点云转换成二维图像后输入网络进行检测，原始点云进行特征提取后实现目标检测。SECOND（sparsely embedded convolutional detection）网络[15] 采用了稀疏卷积的方式提取体素特征，在 3D 卷积的过程中只对包含点云的体素进行卷积，忽略掉那些空的体素格子，从而提高了目标检测的速度。PointPillars 网络[16] 是将 3D 数据转换成 2D 的问题去处理，并提出了伪图像的概念，将点云转换成伪图像输入到现有成熟的二维目标检测网络中进行预测。Yang 等[17] 在 PointPillars 网络的基础上加入了注意力机制，提高了模型提取关键信息的能力，进而提升了检测性能。VoxelNet[18] 首先对原始点云进行体素化，对体素特征学习后输出 3D 目标检测框。Yang 等[19] 提出的 STD（sparse-to-dense）算法通过球形锚框削减锚框的生成数量，能大大减少计算量，但基于欧氏距离的最远点采样算法产生的大量背景点制约了检测效率。3DSSD[20]（point-based 3D single stage object detector）是一种轻量级的 3D 物体检测模型，在点云采样过程中，通过结合欧氏距离和特征距离，对最远点采样算法进行了优化，以解决采样过程中出现包含背景点的问题。PV-RCNN（point-voxel region-based convolutional neural networks）[21] 采用稀疏卷积提取点云的体素特征，通过最远点采样的方式获取点云关键点，将体素特征与点特征进行融合来进行 3D 目标框的回归，虽然在检测精度上有很大提升，但是在大规模点云数据集中检测速度十分慢，原因是点云降采样过程中消耗了大量时间。PointRCNN: 3D object proposal generation and detection from pointcloud 网络[22] 则直接在原始点云上进行特征提取，属于两阶段形式的网络结构，第 1 阶段网络为区域提案网络 (region proposal network, RPN)[23]，采用 PointNet++[24] 作为骨干特征提取网络，通过对点云进行语义分割在所有前景点上生成大量的 3D 提案；第 2 阶段网络为区域卷积网络 (RCNN)[25]，包含 4 个模块：区域池化 (region of interest pooling, RoI pooling)[26] 模块、坐标变换模块、特征提取模块以及精细化回归模块。对第 1 阶段生成的候选目标包围框做精细调整，生成最终结果。PointRCNN 方法虽然检测精度有了显著的提高，但在特征提取阶段使用了最远点采样 (farthest point sampling, FPS)[27] 算法来对点云进行降采样，随着点云输入规模的增加，所消耗的时间也会大幅增加，导致该算法在目标检测的实时性较差。

本文提出了一种两阶段快速目标检测算法 RandLA-RCNN (random sampling and an effective local feature aggregator with RCNN)，利用 PointRCNN 第 2 阶段的 RCNN 网络对第 1 阶段生成的提案框进行精细回归。在第 1 阶段，通过随机采样 (random sampling, RS) 的方式来对点云进行降采样，采样方式不受点云规模的限制，在大规模点云场景中，能够显著提升网络的检测速度。同时，为了在随机采样过程中尽可能保留复杂的局部点云特征，从而提高检测精度，在本文提出了局部特征聚合 (local feature aggregation, LFA) 模块，并利用注意力机制，聚合相邻点云的特征，为了尽可能地保留更多原始点云中的几何信息，提出了扩张残差模块。

1　PointRCNN 网络结构

PointRCNN 是直接在原始点云上进行特征提取并生成 3D 包围框的两阶段检测框架。检测时，第 1 阶段网络首先生成大量的 3D 提案，第 2 阶段对这些提案进行细化。第 1 阶段为 RPN 网络，采用 PointNet++的分割网络提取原始点云的特征，使用多尺度特征融合 (multi-scale grouping) 方法在原始点云上生成逐点的特征向量。同时，对点云进行密集采样，生成大量的候选 3D 框，并使用一个子网来评估每个候选框是否包含物体，并将得分高的候选框保留下来，最终生成高质量的 3D 提案框。PointRCNN 的第 2 阶段网络主要用于对第 1 阶段生成的 3D 提案边界框进行精细化调整。首先将第 1 阶段学习到的点特征汇集到每个 3D 边界框提案中，以便进行后续处理。然后将每个 3D 边界框转换为规范化坐标系中的坐标，并将其与第 1 阶段生成的分割掩码和点特征相结合，以学习调整相对坐标。最后，使用一个多层感知器来预测每个 3D 提案边界框的最终位

置和大小。

同时, 为了帮助模型更好地学习 3D 检测任务, 提高模型的性能和鲁棒性, 在 PointRCNN 中, 使用了多种不同的损失函数来训练模型。其中, 第 1 阶段的损失包含:

1) 对该帧中所有点云前景点分类损失。由于在一帧点云中属于前景点的数量差异较大, 因此使用了 Focal Loss[28] 来解决类不平衡问题, 该部分损失函数为

$$L_{\text{focal}}(p_t) = -\alpha_t(1-p_t)^{\gamma}\log(p_t)$$
$$p_t = \begin{cases} p, & p \in R_{\text{F}} \\ 1-p, & p \in R_{\text{B}} \end{cases} \tag{1}$$

式中: R_{F}为前景点的集合, R_{B}为背景点的集合。

2) 前景点回归损失使用了 smooth L1[29] 损失函数来计算前景点与 ground-truth 之间的损失。第 2 阶段的损失包含前景点 RoI 的置信度损失和前景点 RoI 边界框回归损失, 通过这两个部分的损失函数, 模型可以更准确地检测和定位目标。

2　RandLA-RCNN 模型

本文针对 PointRCNN 第一阶段 RPN 网络中的骨干特征提取网络进行改进, 这是因为在 PointRCNN 网络检测目标时, 模型在特征提取阶段花费时间占总时间的近 50%, 对其改进以提高第 1 阶段生成 3D 候选框的速度和准确性。RandLA-RCNN 网络的第 1 阶段的结构包括骨干特征提取网络和 3D 包围框生成模块。其中, 骨干特征提取网络结构上分为编码层与解码层。编码层由随机采样 RS 模块与局部特征聚合 (LFA) 模块组成, LFA 模块是为了在使用随机采样算法的同时, 将近邻点的局部特征也尽可能保留, 通过 LFA 来编码近邻点的信息, 同时扩充中心点的感受野, 减少特征信息的丢失。LFA 包含 3 个子模块, 分别是局部空间编码 (local spatial encoding, LSE) 模块、注意力池化 (attention pooling, AP) 模块以及扩张残差模块 (dilated residual block, DRB)。解码层由特征传播 (feature propagation, FP) 层以及多层感知机 (multilayer perception, MLP) 组成, 对点云进行上采样, 将点云恢复到输入时的分辨率。网络的第 2 阶段 RCNN 由 RoI pooling 层、特征融合模块及 3 个 SA(set abstraction) 模块组成。网络的整体结构如图 1 所示。

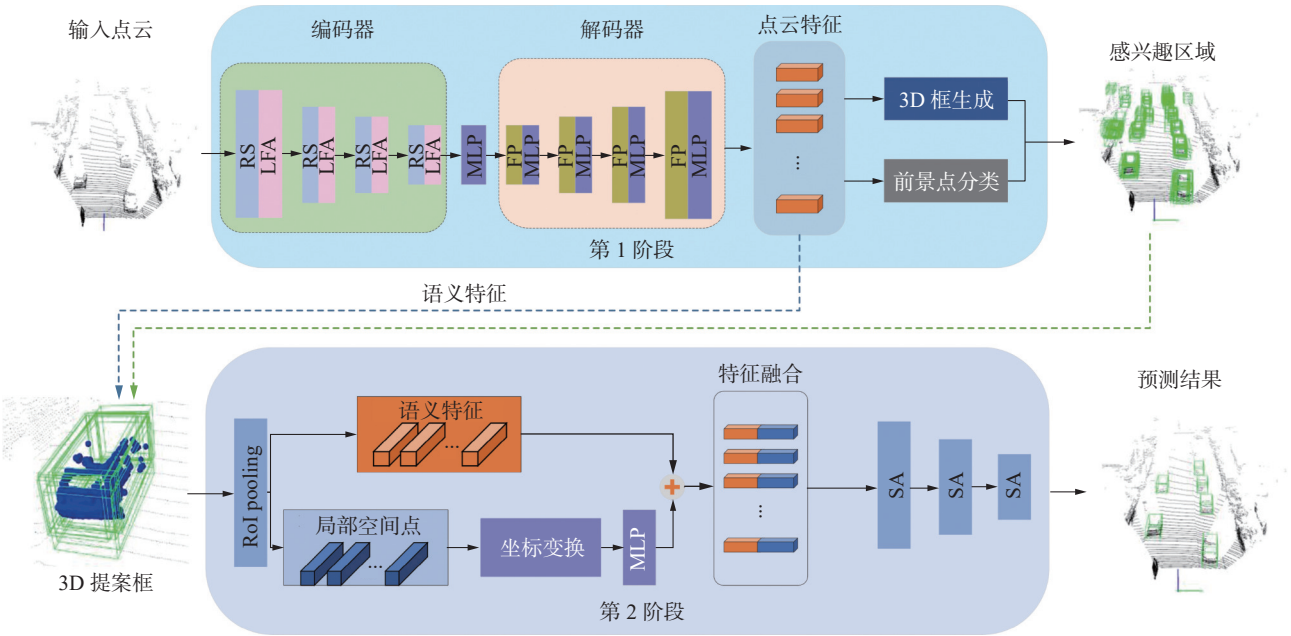


图 1　RandLA-RCNN 整体框架

Fig. 1　Framework diagram of RandLA-RCNN

2.1　点云采样策略

对于大场景的目标检测, 通常涉及数百万甚至更多点云数据。然而, 目前并没有比较标准的采样策略适用于大规模点云场景, 因此高效的采样算法是实现快速检测的必要条件。通过高效的采样算法, 本文模型实现从原始数据中提取关键信息, 以最小化计算和内存需求, 并且能够在保证检测准确率的同时, 实现快速检测。从算法实现角度来说, 现有的点云采样方法可以大致分为启发式和生成式的算法。

2.1.1　启发式点云采样算法

1) 最远点采样

为了从具有N个点的大规模点云中采样k个点, FPS 会对度量空间$p_1, p_2, \cdots, p_k$重新排序, 使

得每个p_k是距离前$k-1$个点最远的点。FPS 虽然能够很好地覆盖整个点集，但是其计算复杂度为$O(N^2)$。

2) 逆密度采样[30](inverse density importance sampling, IDIS)

要从个N点中采样 k 个点，IDIS 根据密度对输入的N个点重新进行排序，然后选择前 k 个点，因此其时间复杂度为 $O(N)$，相较于 FPS 效率有所提高，但在实时系统中仍太慢。

3) 随机采样

随机采样算法是一种简单的点云下采样方法，通过设置采样率r，即从原始点云中保留的点占原始点云总点数N的比例，通常情况下，r的取值范围为$(0,1]$。根据r计算需要保留的点的数量$n=rN$，从N个点中均匀地选择n个点作为采样后的点云，时间复杂度是 $O(1)$。无论点云规模大小如何，与 FPS 和 IDIS 相比，随机采样都具有更高的计算效率。

2.1.2　生成式点云采样算法

1) 基于生成器的采样 (generator-based sampling, GS)[31]

GS 通过生成器生成一个小的点集来近似表示原始的大点集。在推理阶段将生成的子集与原始集合进行匹配时，常要用到 FPS 算法从而产生额外的计算。

2) 基于连续松弛的采样算法 (cont-inuous relaxation sampling, CRS)[32]

CRS 方法使用重参数技巧将采样操作放宽到连续域来进行端到端训练。每个采样点都是基于点云的加权和来学习的。在对所有新数据进行采样时，会产生一个较大的权重矩阵点与单程矩阵乘法同时进行，导致巨大的内存开销。

3) 基于策略梯度的采样方法 (policy gradient sampling, PGS)[33]

PGS 采样的过程近似一个马尔可夫决策过程，通过学习每个点的概率分布来对点云进行降采样，但是当点云规模较大时，得到的不同点之间概率分布差异也很大，不利于网络的收敛。

总的来说，针对现有点云采样算法，FPS、IDIS 和 GS 等存在计算成本过高、不能应用于大规模点云等问题；PGS 算法难以学习，而 CRS 算法具有巨大的内存开销，同样不适合处理大规模点云。相比之下，随机采样算法具有以下两个优点：1) 由于不受点云输入规模限制，它具有高效的计算效率；2) 它不需要额外的存储器进行计算。因此，随机采样是处理大规模点云最合适的方法。

2.2　局部空间编码模块

局部空间编码模块对输入点云中每个近邻点的空间位置特征进行编码，将编码后的特征作为输入点的特征向量，使得在特征提取过程中不同点之间可以相互感知它们之间的相对空间位置，从而能够使 LSE 模块学习复杂的局部结构。具体而言，该模块包括以下步骤:

1) 搜索相邻点

每一个点的局部特征进行编码前，采用基于 K-d 树[34] 的 K 近邻搜索算法搜索每个局部区域内的点云以提高效率。对于三维数据而言，所构造的 K-d 树如图 2 所示。

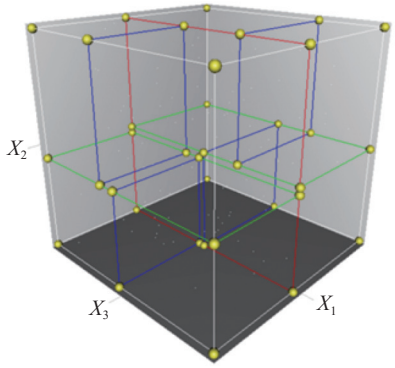


图 2　三维 K-d 树示意

Fig. 2　Schematic diagram of 3D K-d tree

2) 近邻特征嵌入

为了利于网络学习局部特征，获得较好的性能，该模块将对每个中心点周围的 k 个近邻点的语义特征与几何特征进行编码。编码的特征包括中心点坐标、近邻点坐标、中心点与近邻点的相对坐标、以及近邻点到中心点的欧氏距离。编码公式为

$$r_i^k = \mathrm{MLP}\left(p_i \oplus p_i^k \oplus \left(p_i - p_i^k\right) \oplus \left\|p_i - p_i^k\right\|\right) \tag{2}$$

式中：p_i为某个中心点的三维坐标(x,y,z)；p_i^k为p_i周围k个近邻点的三维坐标信息；$\oplus$是串联操作；$\|\cdot\|$是计算两个点之间的欧氏距离；r_i^k是每个点对其周围点编码后的特征，输出特征的维度为$k\times d$。

局部空间编码 (LSE) 模块整体的网络结构如图 3 所示。原始点云首先经过 KNN(k-nearest neighbor) 算法得到其近邻点信息包括三维坐标信息以及特征信息。近邻点的坐标信息经过位置编码后得到，特征信息经过 MLP 后得到，最后将两者串联得到融合特征，该特征融合了中心点的局部特征，将作为注意力池化模块的输入。

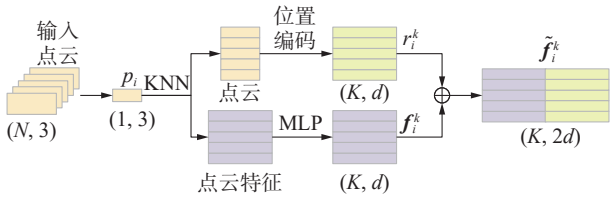


图 3　局部空间编码模块

Fig. 3　Local spatial coding module

2.3　注意力池化模块

注意力池化模块用于聚合每个中心点周围的语义特征, 获取每个局部区域的全局特征。相比原算法中的最大值池化方式, 可能会存在导致大部分信息丢失的风险, 注意力池化机制能获得每个通道的注意力得分, 强化对输出起主导作用通道的权重, 弱化那些对输出影响较小通道的权重, 从而以加权求和方式聚合不同通道的特征信息。该模块详细的结构如图 4 所示。

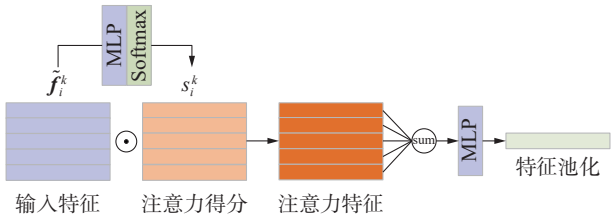


图 4　注意力池化模块

Fig. 4　Attention pooling module

对于某个点 p_i 所对应领域内的特征 $\tilde{\boldsymbol{f}}_i^k$, 设计了一个注意力函数 $g(\cdot)$ 在训练过程中自动去学习每个通道的注意力分数 s_i^k, 在网络中这个函数是通过一个共享权值的 MLP 同时再加入 Softmax 层来实现。

$$s_i^k = g(\boldsymbol{f}_i^k, \boldsymbol{W}) \tag{3}$$

式中 $\boldsymbol{W}$ 表示共享 MLP 中的权值参数。

池化特征的计算公式为

$$\tilde{\boldsymbol{f}}_i = \mathrm{MLP}\left(\sum_{k=1}^{K}(\tilde{\boldsymbol{f}}_i^k \cdot s_i^k)\right) \tag{4}$$

将输入的近邻点特征 $\tilde{\boldsymbol{f}}_i^k$ 与注意力分数 s_i^k, 通过点积运算得到注意力特征, 再将近邻点特征进行累加, 再通过一个 MLP 层得到池化后的全局特征。对于第 i 个点 p_i, 经过 LSE 和 AP 模块, 能够聚合其 k 个最近点的几何特征, 并最终输出信息特征向量 $\tilde{\boldsymbol{f}}_i$。

2.4　扩张残差模块

在点云的特征提取网络中, 随着对点云不断地降采样, 原始空间中的一些点云也随之丢弃掉, 为了尽可能地保留更多原始点云中的语义信息, 需要显著地增加经过降采样后每个中心点的感受野, 使得在当前层的特征输入到下一层中仍然保留原始点云中的一些几何特征, 即使这些点云在前面已经被丢弃掉。扩张残差模块 (dilated residual block, DRB) 采用了多层堆叠以及跳跃连接的网络结构, 在每个 LFA 模块中堆叠了多个 LSE 与 AP 模块, 并采用跳级结构与输入端进行连接, 结构如图 5 所示。

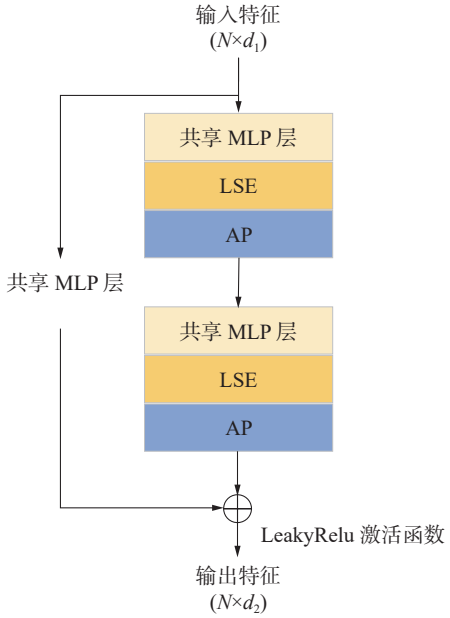


图 5　扩张残差模块结构

Fig. 5　Expansion residual module structure diagram

由图 5 可以看到, 该模块首先对输入的点云特征连续进行了两次局部空间编码与注意力池化操作, 同时使用跳级结构将输入端通过共享权值的多层感知机与输出端进行连接, 末端采用 LeakyRelu 函数作为激活函数得到聚合特征向量。

为了进一步展示该模块在拓展中心点感受野的有效性, 以某个局部区域的点云来进行说明, 具体的效果如图 6 所示。

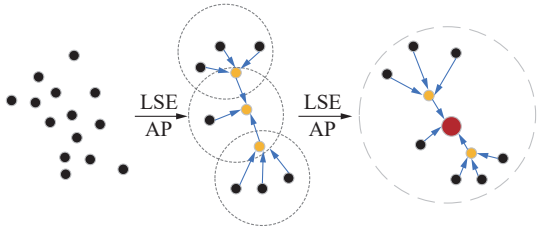


图 6　感受野扩张示意

Fig. 6　Schematic diagram of receptive field expansion

最左侧表示原始的输入点云, 输入端点云经过第 1 次 LSE 与 AP 操作后, 每个中心点即图 6 中间部分的黄点, 此时它包含了以它为中心, 周围 3 个点的特征信息, 设此时它的感受野为 K。经过第 2 次同样操作后, 中心点变为图中最右侧部分中的红点, 其特征向量中包含整个局部区域所有点的特征信息, 此时该中心点的感受野变为

K^2。通过在网络中不断堆叠该模块来不断扩大点云的感受野从而保留更多原始点的特征信息，但是这样会带来计算上的代价，使得模型整体的运算效率变低。本文采用了两级串联的方式，实验结果也表明，在保证计算效率的前提下，进行两次扩张可以保证很好的检测效果。

2.5　损失函数

在 RPN 网络生成候选目标包围框阶段，网络同时要进行两个任务分支，分别是语义分割头的分类损失以及回归头的回归任务损失，考虑到室外点云场景中通常存在正负样本不均衡的情况，以 KITTI 数据集为例，平均一帧的点云场景中大约有 10 万个点，而前景点的数量可能只有几千个，不利于分类任务的训练，因此分割头采用 Focal loss 作为损失函数。回归头部分的回归参数包含中心点坐标的回归、包围框尺寸的回归以及朝向角的回归。对于垂直于地面的方向也就是相机坐标系中的坐标 y 以及包围框尺寸的回归，采用直接回归的方式，损失函数采用 smooth L1，该部分整体的损失函数为

$$L_{\text{res}}^{p}=\sum_{v\in\{y,h,w,l\}}F_{\text{reg}}(\text{res},\widehat{\text{res}}) \tag{5}$$

式中：res表示各参数回归的偏差值，$\widehat{\text{res}}$表示实际标签值。

对于候选框在 x 轴和 z 轴以及朝向角度的回归，采用基于刻度的损失函数，刻度分类任务采用交叉熵损失函数，残差项的回归采用 smooth L1 损失，总体的损失函数为

$$L_{\text{bin}}^{p}=\sum_{u\in\{x,z,\theta\}}\left[F_{\text{cls}}\left(\text{bin}_{u}^{p'},\text{bin}_{u}^{p}\right)+F_{\text{reg}}(\text{res},\widehat{\text{res}})\right] \tag{6}$$

式中：F_{cls}表示交叉熵损失，F_{reg}表示 smooth L1 损失。

因此，第 1 阶段网络中回归头部分的损失函数为

$$L_{\text{res}}=\frac{1}{N_{\text{pos}}}\sum_{p\in\text{pos}}(L_{\text{bin}}^{p}+L_{\text{res}}^{p}) \tag{7}$$

式中：N_{pos}表示前景点的数量，pos 为前景点集合。

3　实验分析

3.1　实验设置

本文在 KITTI 数据集上评估了 RandLA-RCNN 在汽车检测任务中的性能，评估了检测精度、检测速度两个方面。精度评估根据目标是否遮挡、是否被截断以及目标框的边界范围将检测难度划分为简单、中等和困难 3 个等级，具体标准见表 1。

表 1　目标检测困难程度划分标准

Table 1　Difficulty classification criteria for target detection

难度	RGB图像中边界框范围/像素	遮挡类型	截断程度/%
简单	40	无遮挡	15
中等	25	小部分遮挡	30
困难	25	大部分遮挡	50

第 1 阶段网络输入数据为固定采样的 16 384 个点，每层输入的点云数目 (16 384, 4 096, 1 024, 256, 64)，每一层特征通道数设置为 (8, 32, 128, 256, 512)。网络通过 RoI Pooling 层进行区域兴趣的提取，每个候选区域扩大的范围设置为 1.0 m，采样点的数目设置为 512，每个 SA 模块，球查询半径集合设置为 [0.2 m, 0.4 m]，每个球查询空间内采样点的数目均设置为 64。对网络中每个真值框所在的区域扩大 0.2 m，扩大区域内的点云不参与损失函数的计算，标签值设为−1。训练分两阶段，第 1 阶段网络训练的每个批次 Batch size 为 16，训练 200 个 epoch，学习率为 0.002；第 2 阶段训练 RCNN 网络，每个批次 Batch size 为 8，训练 50 个 epoch，学习率 为 0.002，使用 Adam 优化器。第 2 阶段网络其余设置参数及损失函数项与文献 [11] 默认值一致。

3.2　随机采样算法有效性实验

为了评估现有采样方法处理不同规模点云数据的效率，包括 FPS、IDIS、RS、GS、CRS 和 PGS 方法。本文进行了 4 组实验，分别对 10^4、10^5、10^6、10^7 级别规模的点云进行下采样，每组仅保留原始点云规模的 30%，对比每种采样方法的耗费时间，实验结果如图 7 所示。

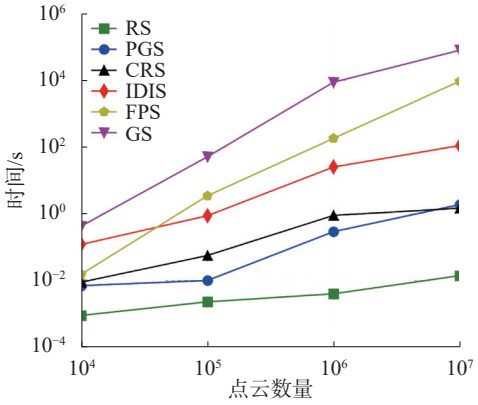


图 7　不同采样方法消耗时间

Fig. 7　Time consumption of different sampling approaches

图 7 比较了不同采样方法处理不同规模数量的点云的总时间。可以看出，对于小规模点云，所有采样方法耗费时间相近，并不会有较大的计算压力，而当点云规模较大时，PGS、CRS、IDIS、FPS、GS 耗费的时间相比 RS 方法多出几个数量

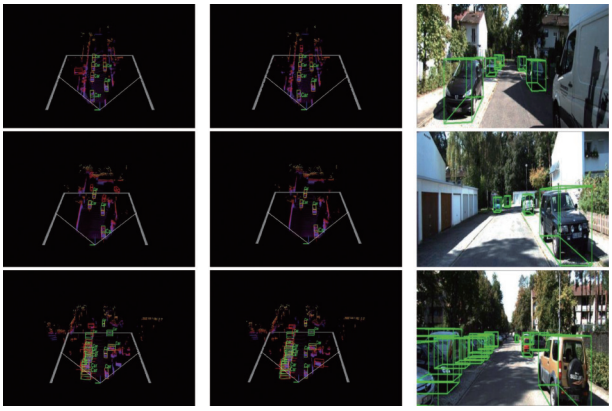
级。因此, 本文在 RandLA-RCNN 中使用随机采样算法具有高效的处理效率。

3.3　汽车和行人目标检测实验

汽车类目标检测的可视化结果如图 8 所示, 鸟瞰图中的可视化结果如图 9 所示。点云可视化结果中绿色的框代表预先标注好的真值框, 红色的框代表 RandLA-RCNN 预测的结果, 鸟瞰图结果中图中蓝色框表示预测框, 绿色框表示真值框。

由图 8、9 可以看到, 在第 1 个场景中, PointRCNN 中有误检的情况, 将车后的树识别为汽车, 改进后的 RandLA-RCNN 没有出现误检的情况, 同时, 对于远处的车辆, 虽然未被标注, 但是 RandLA-RCNN 仍可将该车辆准确地检测到。第 2 个场景中原算法同样存在误检的情况, RandLA-RCNN 将标注的 4 辆车都准确地检测到了, 且没有出现误检的情况。第 3 个场景属于相对复杂的场景, 包含目标数量较多, 对于大部分目标车辆, 两种算法均能准确地检测到。

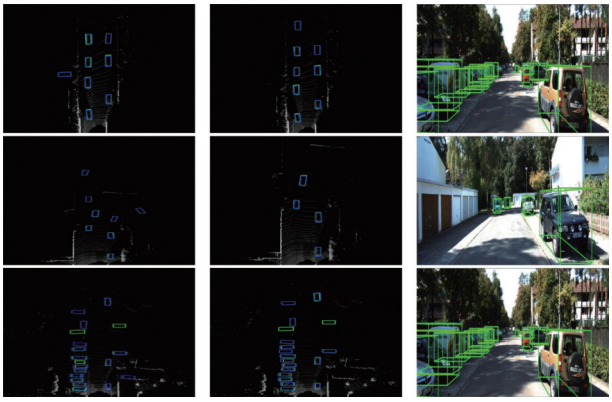
本文对车辆和行人目标交并比 (intersect- ion over union, IoU) 阈值分别设置为 0.7 和 0.5, 进行了不同算法在处理单帧点云所消耗的时间以及在验证集上的检测结果的比较, 结果如表 2 和 3 所示。从表 2 中数据可以看出, 相较于改进前的算法, 本文提出的 RandLA-RCNN 在检测速度上由原来的 0.12s 提升为 0.06 s, 提升到原来的 2 倍。相较于 SECOND 网络, RandLA-RCNN 网络的检测速度稍慢, 这是因为 SECOND 网络是单阶段的网络结构, 在检测速度上要明显优于两阶段的检测网络, 但本文算法在 3 个不同难度的检测精度均优于 SECOND 网络, 而算法设计之初就是期望能够保证高精度的同时, 具有比较快的检测速度。



(a) 原算法　(b) RandLA-RCNN　(c) 点云场景所对应 RGB 图像

图 8　车辆检测在点云中的可视化结果

Fig. 8　Visualization of vehicle detection in point cloud



(a) 原算法　(b) RandLA-RCNN　(c) 点云场景所对应 RGB 图像

图 9　车辆检测在鸟瞰图中的可视化结果

Fig. 9　Visualization of vehicle detection in bird's eye view

表 2　各算法在 KITTI 验证集中车辆和行人目标的 3D 检测结果

Table 2　3D detection results of vehicle targets in the KITTI validation set for each algorithm

算法	时间/ s	车辆(IoU=0.7)/%				行人 (IoU=0.5)/%			
		简单	中等	困难	mAP	简单	中等	困难	mAP
MV3D[35]	0.36	71.29	62.68	56.56	63.51	—	—	—	—
F-Point Net	0.17	83.76	70.92	63.56	72.75	—	—	—	—
Voxel Net	0.23	81.98	65.46	62.85	70.10	—	—	—	—
SECONED	**0.03**	87.43	76.48	69.10	77.67	—	—	—	—
PointRCNN	0.12	87.92	**77.69**	**76.43**	80.68	50.22	42.01	39.63	43.95
RandLA-RCNN	0.06	**88.64**	77.62	76.24	**80.83**	**50.31**	**42.03**	39.59	**43.98**

表 3　各算法在 KITTI 验证集中车辆和行人目标 BEV 的检测结果

Table 3　BEV detection results of vehicle target in the KITTI validation set for each algorithm

算法	时间/s	车辆(IoU=0.7)/%				行人(IoU=0.5)/%			
		简单	中等	困难	mAP	简单	中等	困难	mAP
MV3D	0.36 s	86.55	78.10	76.67	80.44	—	—	—	—
F-Point Net	0.17 s	88.16	84.02	76.44	84.45	—	—	—	—

续表 3

算法	时间/s	车辆(IoU=0.7)/%				行人(IoU=0.5)/%			
		简单	中等	困难	mAP	简单	中等	困难	mAP
Voxel Net	0.23 s	89.60	84.81	78.57	84.33	—	—	—	—
SECONED	**0.03 s**	89.96	87.07	79.66	85.56	—	—	—	—
PointRCNN	0.12 s	91.21	86.89	82.51	**86.87**	56.62	49.43	44.78	50.28
RandLA-RCNN	0.06 s	**91.33**	86.78	82.41	86.84	**56.78**	49.39	44.73	**50.30**

在检测精度方面，RandLA-RCNN 与 PointRCNN 相比，平均精准率 (average precision) 基本一致，在车辆类简单目标上的检测精度提升了 0.72 百分点，在中等复杂目标的检测精度略有降低，原因可能是困难目标样本本身的点云数据较少，随机采样的过程中容易丢失掉，导致该类目标的检测效果略有下降，但是在 3 个难度的平均精度 (mean average precision, mAP) 提升了 0.15 百分点。在行人目标方面，对简单以及中等目标的检测精度上分别提升了 0.09 和 0.02 百分点，而对困难物体的检测精度下降了 0.04 百分点，平均精度提升了 0.03 百分点。在点云鸟瞰图的检测结果中，检测精度与 PointRCNN 基本一致。总体来看，考虑检测精度和检测速度的同时，本文所提算法均优于其他算法。

3.4　消融实验

本文进行了时间消耗的对照实验，计算了两种算法在不同模块的消耗时间，结果如表 4。在特征提取阶段，PointRCNN 耗时 60 ms，占总时间近 50%，而本文改进的 RandLA-RCNN 的推理速度提升到 20 倍。在其余模块中，两种算法消耗时间基本相同。综合来看，改进后的 RandLA-RCNN 检测速度提升近两倍，达到 16 f/s。

表 4　改进后算法与原算法各模块检测时间对比
Table 4　Comparison of the detection time of each module between the improved algorithm and the original algorithm　ms

算法	骨干特征提取网络	ROI生成	第2阶段网络	总时间
PointRCNN	60	42	22	124
RandLA-RCNN	**3**	**39**	22	**64**

本文还对 LSE 模块以及 DRB 的有效性进行了验证。对局部空间编码模块分别比较了以下情况对检测精度的影响：只编码中心点坐标 p_i、只编码近邻点坐标 p_i^k、编码中心点以及近邻点坐标 (p_i, p_i^k)、编码中心点近邻点坐标以及中心点与近邻点的欧几里得距离 $\left(p_i, p_i^k, \left\|p_i - p_i^k\right\|\right)$、编码中心点近邻点坐标以及相对位置 $(p_i, p_i^k, p_i - p_i^k)$，实验结果如表 5 所示。

表 5　LSE 模块对比实验结果
Table 5　LSE module comparison experiment results　%

编码的特征信息	mAP
(p_i)	74.32
(p_i^k)	75.43
(p_i, p_i^k)	75.49
$\left(p_i, p_i^k, \left\|p_i - p_i^k\right\|\right)$	76.86
$(p_i, p_i^k, p_i - p_i^k)$	79.92
$\left(p_i, p_i^k, p_i - p_i^k, \left\|p_i - p_i^k\right\|\right)$	**80.83**

同时，RandLA-RCNN 中用了两次局部空间编码与注意力池化作为一个标准残差扩张模块，为了验证残差模块中局部特征聚合单元的数量对检测结果的影响，对比了只进行 1 次特征聚合与进行 3 次特征聚合的实验结果，结果如表 6。

表 6　DRB 模块对比实验结果
Table 6　DRB module comparison experiment results　%

残差扩张模块DRB	简单
1次特征融合	78.01
2次特征融合(标准)	**80.83**
3次特征融合	80.28

通过表 6 中的结果可以分析，只进行 1 次局部特征聚合，检测精度降低，而进行 3 次特征聚合并未按预期提升检测精度，这因为采用多次特征聚合会使模型参数增加，导致过拟合，因此采用两次特征聚合更适用当前的检测场景。

4　结束语

针对当前两阶段点云 3D 目标检测网络检测速度较慢的问题，提出了一种基于随机采样和局部特征聚合的目标检测算法，改进对车辆检测精度较高的 PointRCNN 网络。本文算法采用随机采样的方法，提升对目标的检测速度，通过局部特征聚合模块来扩大中心点的感受野，保留更多原始点云的特征信息，从而在随机采样过程中尽可能保留近邻点的局部特征信息，提高对目标的检测精度。实验表明，提出的 RandLA-RCNN 网络在检测精度保持较高水平的同时，显著提升了

对车辆目标的检测速度，为高效、准确检测 3D 目标提供了一种新思路。

参考文献：

[1] HUANG Keli, SHI Botian, LI Xiang, et al. Multi-modal sensor fusion for auto driving perception: a survey[EB/OL]. (2022−02−06)[2023−11−13]. https://arxiv.org/abs/2202.02703.

[2] 刘通, 高思洁, 聂为之. 基于多模态信息融合的多目标检测算法[J]. 激光与光电子学进展, 2022, 59(8): 339−348.
LIU Tong, GAO Sijie, NIE Weizhi. Multitarget detection algorithm based on multimodal information fusion[J]. Laser & optoelectronics progress, 2022, 59(8): 339−348.

[3] SONG Ziying, LIU Lin, JIA Feiyang, et al. Robustness-aware 3D object detection in autonomous driving: a review and outlook[EB/OL]. (2024−01−12)[2024−08−02]. http://arxiv.org/abs/2401.06542v3.

[4] VORA S, LANG A H, HELOU B, et al. PointPainting: sequential fusion for 3D object detection[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020: 4603−4611.

[5] WANG Chunwei, MA Chao, ZHU Ming, et al. PointAugmenting: cross-modal augmentation for 3D object detection[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021: 11789−11798.

[6] XU Shaoqing, ZHOU Dingfu, FANG Jin, et al. FusionPainting: multimodal fusion with adaptive attention for 3D object detection[C]//2021 IEEE International Intelligent Transportation Systems Conference. Indianapolis: IEEE, 2021: 3047−3054.

[7] BAI Xuyang, HU Zeyu, ZHU Xinge, et al. TransFusion: robust LiDAR-camera fusion for 3D object detection with transformers[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022: 1080−1089.

[8] LIANG Tingting, XIE Hongwei, YU Kaicheng, et al. Bevfusion: A simple and robust lidar-camera fusion framework[J]. Advances in neural information processing systems, 2022, 35: 10421−10434.

[9] LI Yingwei, YU A W, MENG Tianjian, et al. DeepFusion: lidar-camera deep fusion for multi-modal 3D object detection[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022: 17161−17170.

[10] HU Haotian, WANG Fanyi, SU Jingwen, et al. EA-BEV: edge-aware bird's-eye-view projector for 3D object detection[EB/OL]. (2023−03−31)[2023−11−13]. https://arxiv.org/abs/2303.17895.

[11] YAN Junjie, LIU Yingfei, SUN Jianjian, et al. Cross modal transformer via coordinates encoding for 3D object dectection[EB/OL]. (2023−01−03)[2023−11−13]. https://arxiv.org/abs/2301.01283.

[12] WANG Haiyang, TANG Hao, SHI Shaoshuai, et al. UniTR: a unified and efficient multi-modal transformer for bird's-eye-view representation[C]//2023 IEEE/CVF International Conference on Computer Vision. Paris: IEEE, 2023: 6792−6802.

[13] 张新钰, 邹镇洪, 李志伟, 等. 面向自动驾驶目标检测的深度多模态融合技术[J]. 智能系统学报, 2020, 15(4): 758−771.
ZHANG Xinyu, ZOU Zhenhong, LI Zhiwei, et al. Deep multi-modal fusion in object detection for autonomous driving[J]. CAAI transactions on intelligent systems, 2020, 15(4): 758−771.

[14] 鲁斌, 杨振宇, 孙洋, 等. 基于多通道交叉注意力融合的三维目标检测算法[J]. 智能系统学报, 2024, 19(4): 885−897.
LU Bin, YANG Zhenyu, SUN Yang, et al. 3D object detection algorithm with multi-channel cross attention fusion[J]. CAAI transactions on intelligent systems, 2024, 19(4): 885−897.

[15] YAN Yan, MAO Yuxing, LI Bo. SECOND: sparsely embedded convolutional detection[J]. Sensors, 2018, 18(10): 3337.

[16] LANG A H, VORA S, CAESAR H, et al. PointPillars: fast encoders for object detection from point clouds[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 12689−12697.

[17] YANG Zetong, ZHOU Yin, CHEN Zhifeng, et al. 3D-MAN: 3D multi-frame attention network for object detection[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021: 1863−1872.

[18] ZHOU Yin, TUZEL O. VoxelNet: end-to-end learning for point cloud based 3D object detection[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018.

[19] YANG Zetong, SUN Yanan, LIU Shu, et al. STD: sparse-to-dense 3D object detector for point cloud[C]//2019 IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019: 1951−1960.

[20] YANG Zetong, SUN Yanan, LIU Shu, et al. 3DSSD: point-based 3D single stage object detector[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020: 11037−11045.

[21] SHI Shaoshuai, GUO Chaoxu, JIANG Li, et al. PV-RCNN: point-voxel feature set abstraction for 3D object detection[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020: 10526−10535.

[22] SHI Shaoshuai, WANG Xiaogang, LI Hongsheng. PointRCNN: 3D object proposal generation and detection from point cloud[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 770−779.

[23] LI Bo, YAN Junjie, WU Wei, et al. High performance visual tracking with siamese region proposal network[C]// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 8971−8980.

[24] QI C R, YI Li, SU Hao, et al. PointNet++: deep hierarchical feature learning on point sets in a metric space[EB/OL]. (2017−06−07)[2023−11−13]. http://arxiv.org/abs/1706.02413v1.

[25] JIANG Yingying, ZHU Xiangyu, WANG Xiaobing, et al. R2CNN: rotational region CNN for orientation robust scene text detection[EB/OL]. (2017−06−29)[2023−11−13]. https://arxiv.org/abs/1706.09579.

[26] HOU Yi, ZHANG Hong, ZHOU Shilin, et al. Efficient ConvNet feature extraction with multiple RoI pooling for landmark-based visual localization of autonomous vehicles [J]. Mobile information systems, 2017: 8104386.

[27] LI Yangyan, BU Rui, SUN Mingchao, et al. Pointcnn: convolution on X-transformed points[J]. Advances in neural information processing systems, 2018, 31: 820−830.

[28] LIN T Y, GOYAL P, GIRSHICK R, et al. Focal loss for dense object detection[C]//2017 IEEE International Conference on Computer Vision. Venice: IEEE, 2017: 2999−3007.

[29] WANG Hua, NIE Feiping, HUANG Heng. Robust distance metric learning via simultaneous l1-norm minimization and maximization[C]//Proceedings of the 31st International Conference on Machine Learning. Beijing: PMLR, 2014: 1836−1844.

[30] GROH F, WIESCHOLLEK P, LENSCH H P A. Flex-convolution[C]//Asian Conference on Computer Vision. Cham: Springer, 2018: 105−122.

[31] DOVRAT O, LANG I, AVIDAN S. Learning to sample [C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 2755−2764.

[32] YANG Jiancheng, ZHANG Qiang, NI Bingbing, et al. Modeling point clouds with self-attention and gumbel subset sampling[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 3318−3327.

[33] XU K, BA J L, KIROS R, et al. Show, attend and tell: Neural image caption generation with visual attention[C]// Proceedings of the 32nd International Conference on International Conference on Machine Learning. Lille: ICML, 2015: 2048−2057.

[34] BROWN R A. Building a balanced k-d tree in O(kn log n) time[EB/OL]. (2014−10−20)[2023−11−13]. http://arxiv.org/abs/1410.5420v46.

[35] CHEGN Xiaozhi, MA Huimin, WAN Ji, et al. Multi-view 3d object detection network for autonomous driving[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017: 1907−1915.

作者简介：

陆军，教授，博士生导师，博士，主要研究方向为计算机视觉、机器感知、机械臂控制。科技部科技型中小企业创新基金项目评审专家，国家自然科学基金同行评议专家。编写著作 5 部，发表学术论文 80 余篇。E-mail：lujun0260@sina.com。

鲁林超，硕士，主要研究方向为三维目标检测、计算机视觉。E-mail：llczsr@163.com。

翟晓阳，硕士，主要研究方向为三维目标检测、计算机视觉。E-mail：769987461@qq.com。