联合多视角互投影融合的三维目标检测方法

赵亚男[1]，王显才[1]，高利[1]，刘语佳[2]，戴钰[1]
（1. 北京理工大学 机械与车辆学院，北京　100081；2. 天津航海仪器研究所，天津　300130）

摘　要：针对当前智能车辆目标检测时缺乏多传感器目标区域特征融合问题，提出了一种基于多模态信息融合的三维目标检测方法. 利用图像视图、激光雷达点云鸟瞰图作为输入，通过改进 AVOD 深度学习网络算法，对目标检测进行优化；加入多视角联合损失函数，防止网络图像分支退化. 提出图像与激光雷达点云双视角互投影融合方法，强化数据空间关联，进行特征融合. 实验结果表明，改进后的 AVOD-MPF 网络在保留 AVOD 网络对车辆目标检测优势的同时，提高了对小尺度目标的检测精度，实现了特征级和决策级融合的三维目标检测.

关键词：智能车辆；多视角；三维目标检测；互投影融合；AVOD 网络
中图分类号：TP391　　**文献标志码：**A　　**文章编号：**1001-0645(2022)12-1273-10
DOI: 10.15918/j.tbit1001-0645.2021.332

3D Target Detection Method Combined with Multi-View Mutual Projection Fusion

ZHAO Yanan[1]，　WANG Xiancai[1]，　GAO Li[1]，　LIU Yujia[2]，　DAI Yu[1]
（1. School of Mechanical Engineering, Beijing Institute of Technology, Beijing 100081, China;
2. Tianjin Navigation Instrument Research Institute, Tianjin 300130, China）

Abstract: Aiming at the lack of feature fusion of multi-sensor target regions in the current target detection of intelligent vehicles, a three-dimensional target detection method was proposed based on multi-modal information fusion. Firstly, taking the image view and aerial view of lidar point cloud as input, the target detection was optimized by an improved AVOD deep learning network algorithm. And then, a multi-angle joint loss function was inducted to prevent the branch network image degradation. Finally, a dual-view image and the lidar point cloud projected mutual fusion method was presented to enhance data spatial correlation and to carry out feature fusion. The experimental results show that the improved AVOD-MPF network can improve the detection accuracy of small-scale targets while retaining the advantages of the AVOD network for vehicle target detection, and achieve 3D target detection with feature-level and decision-level fusion.

Key words: intelligent vehicle; multi-angle; three-dimensional target detection; mutual projection fusion; AVOD network

环境感知是智能车辆关键性技术，通过车载传感器获取周围环境信息，然后对信息做出分析反馈，主要任务之一是进行目标检测. 目标检测主要包含两个阶段：感兴趣区域生成阶段；三维包围框回归阶段，主要进行感兴趣区域提炼，划分目标类别及尺寸[1－2].

目前普遍采用多传感器融合的方式进行目标检测，主要有决策级融合与特征级融合两种方式. F-pointNet[3] 是一种典型的基于决策级融合二维区域建议网络，从图像中提取二维感兴趣区域，投影到三维激光雷达点云中获得三维视锥，之后输送到 PointNet 为基础的目标检测网络，进行三维包围框预测. ASVADI[4] 等利用 YOLOv3 网络，将 RGB 图像和点云深度图与反射强度图结合，进行特征层融合. 姚钺等[5]

收稿日期：2021－11－30
基金项目：国家重点研发计划项目(2018YFB0105205-02, 2017YFC0804808, 2017YFC0804803)
作者简介：赵亚男(1972—)，女，副教授，博士后，E-mail: zyn@bit.edu.cn.
通信作者：王显才(1996—)，男，硕士生，E-mail: 1737305157@qq.com.

利用 Pointnet++提取特征并进行目标分类与包围框回归. VORA 等[6] 提出了 PointPainting 特征融合模式，将激光雷达点云数据投影到图像坐标，将图像上的分类关联到激光雷达点云，被赋予分类权重的激光雷达点云作为不同目标检测网络（如 Point RCNN[7], Second[8], VoxelNet[9], PointPillar[10] 等）的输入，增强分类性能.CHEN 等[11] 提出一种多视图多模态融合模型-MV3D 网络，在激光雷达点云鸟瞰图上提取三维区域建议，投影到激光雷达前视图与图像平面，提取感兴趣特征，通过多次融合回归得到目标类别与三维包围框. SINDAGI 等[12] 提出了 MVX-Net 网络，将激光雷达点云数据投影到图像特征空间并体素化，非空的体素投影到图像特征空间后获得体素特征，输入到 VoxelNet 网络中. SONG 等[13] 将图像颜色信息扩展到体素通道，引入 3D 离散卷积神经网络改进目标检测网络.

基于多传感器的目标检测方法虽然可以改善单一种类传感器局限[14]，但是当前检测方法多集中于决策级融合，并且不同传感器信息分支在训练中容易退化，导致信息不能完全利用，并且对小尺度目标检测精度都有待提升.

文中提出一种基于图像和激光雷达点云数据的联合多视角目标检测方法，利用包含特征级和决策级融合的 AVOD 网络，通过对多视角信息标注损失函数的优化，避免图像分支网络在训练时退化. 通过互投影池化层对不同模态数据进行特征级融合，对网络目标检测性能有所提高，尤其对小尺度如行人和骑车人目标检测精度提高显著.

1 联合多视角目标检测网络

网络使用来自激光雷达点云鸟瞰图数据和相机前视图 RGB 图像数据，在两个阶段均进行融合操作，融合与检测在网络内不断交替进行，是包含特征级和决策级融合的深度融合网络. 文中整体框架如图 1 所示.

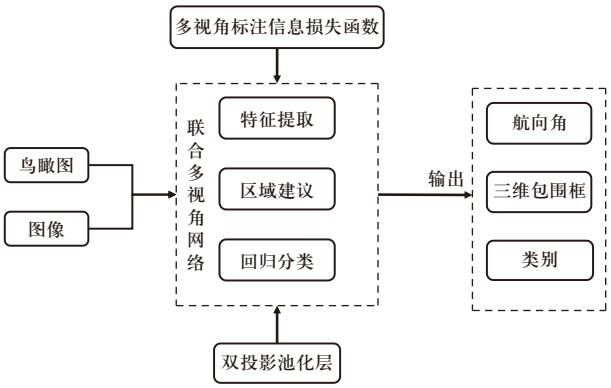


图 1 三维目标检测网络框架

Fig. 1 3D object detection network framework

联合多视角目标检测网络（AVOD）利用激光雷达和图像的信息进行融合，包括两个阶段：初始预测和检测回归. 分别包含数据预处理、特征提取、候选框推理、候选框融合、候选框筛选；候选框投影、特征融合、推理航向角、三维包围框尺寸、目标类别. 其框架如图 2 所示.

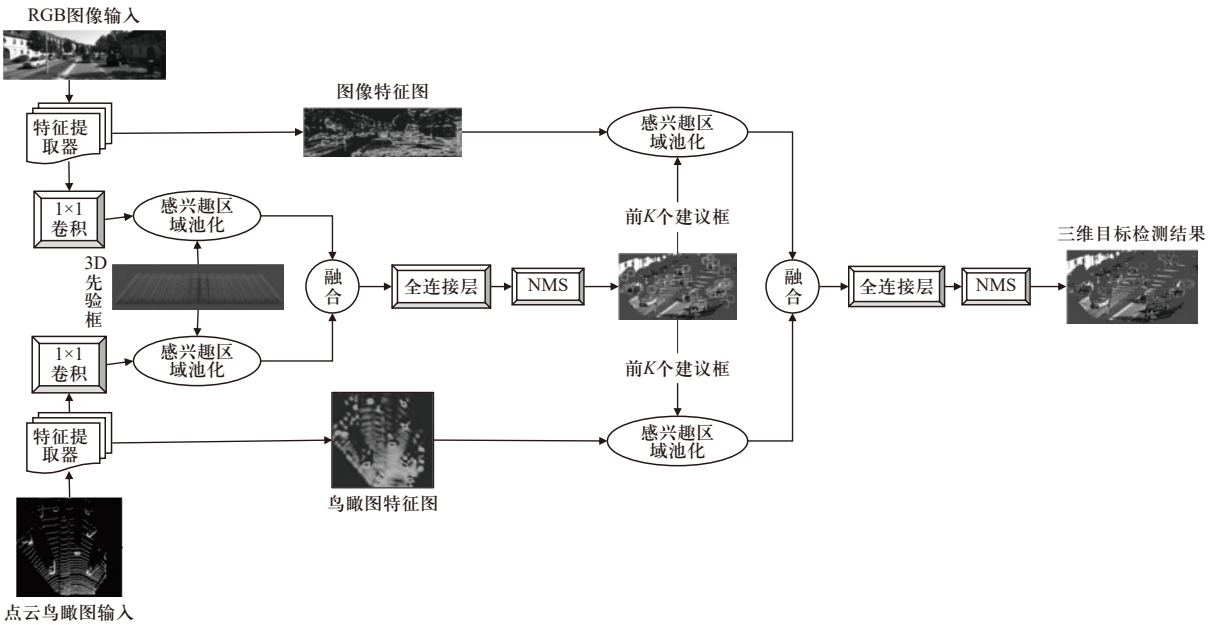


图 2 联合多视角目标检测深度融合网络系统架构

Fig. 2 The architecture of deep fusion network system for joint multi-view target detection

1.1　特征提取与区域建议网络

AVOD 网络第一阶段由特征提取网络和区域建议网络组成, 初步完成精度较低、召回率较高候选框的生成, 尽量避免漏检.

1.1.1　多尺度特征提取网络

特征提取网络综合了 VGG-16 网络和特征金字塔网络(FPN)结构[15]. 选定原型为 VGG-16 的特征提取器, 将各自模态信息进行特征提取. 网络层数加深后, 仅 VGG-16 特征提取器得到的特征图分辨率会越来越低, 对于小尺度目标而言, 其特征随着不断的下采样而丢失, 使得网络丢失小尺度目标检测能力, 因此引入特征金字塔以解决多尺度目标特征提取问题. 特征金字塔的结构可以获得所需的深层次网络蕴含的语义信息, 同时保留浅层网络蕴含的原始细节信息.

特征金字塔是编码器加上解码器结构, 编码由 VGG-16 完成, 解码是一个通过逐步分层恢复分辨率的过程, 利用反卷积对上一特征图进行上采样, 保证提升分辨率的同时保留高层语义信息.

1.1.2　空间 3D 候选框生成

首先进行空间 3D 先验框的构建. 先验框是一系列被预设好具有不同尺寸、宽高比的框, 旨于尽快对目标定位, 提高召回率, 引入 3D 先验框对激光雷达点云鸟瞰图进行目标先验框的处理. 通过 K-means 聚类得到样本先验框尺寸, 以轴对齐的方式编码获得 6 个参数 $(c_x,c_y,c_z,d_x,d_y,d_z)$ 表示的先验框, 其中 (c_x,c_y,c_z) 为中心点坐标, (d_x,d_y,d_z) 为先验框各维度尺寸. (c_x,c_z) 在 x、z 平面上采样, 间隔为 0.5 m, c_y 取决于传感器与地面垂向距离.

该网络将车辆聚类为两种尺寸, 将行人与骑车人聚类为一种尺寸, 每种尺寸设置两个角度(0°、90°)的位姿, 粗略表示目标的不同航向角. 需要筛选并移除稀疏激光雷达点云得到的空白先验框, 保证每帧数据 10 k ~ 100 k 个有效框. 根据空间 3D 先验框获取不同模态下的特征图区域, 并将有效框通过坐标转换分别投影到鸟瞰图和 RGB 图像上, 经过裁剪及双边滤波调整分辨率为 3×3, 便于进行区域融合.

对于复杂场景而言, 先验框数量可能保留到 100 k, 需要使用 1×1 卷积核对特征图降维以减轻后续网络运算负担, 其作用体现在保留不同维度信息的同时大幅减少运算量, 同时实现不同模态特征图跨通道特征级融合. 拼接不同模态的同一先验框中特征图进行拼接, 再利用 1×1 卷积对新张量进行卷积运算.

$$f_o = \sigma\left(\sum_{i=0}^{256}\omega_i f_i + b\right) \tag{1}$$

式中: ω_i 为待学习的权重; f_o 为特征图每个通道包含像素值; b 为偏移量.

之后将融合后的特征图送到两个全连接层, 进行前背景推理和三维包围框回归, 得到规范化后的参数 $(\Delta t_x,\Delta t_y,\Delta t_z,\Delta d_x,\Delta d_y,\Delta d_z)$, 其中 $(\Delta t_x,\Delta t_y,\Delta t_z)$ 是规范化后的中心偏移量, $(\Delta d_x,\Delta d_y,\Delta d_z)$ 为规范化后的尺寸缩放量.

$$\Delta t = (t_g - t_{anchor})/d_{anchor} \tag{2}$$

$$\Delta d = \frac{d_g}{d_{anchor}} \tag{3}$$

损失函数计算采用 Smooth L1 与交叉熵函数的多任务策略, 对三维包围框和目标二元分类分别进行计算, 损失函数为

$$\boldsymbol{L}(\{\boldsymbol{p_i}\},\ \{\boldsymbol{t_i}\}) = \frac{1}{N_{obj}}\sum \boldsymbol{L}_{obj}(p_i,p_i^*) + \frac{\lambda}{N_{reg}}\sum p_i^*\boldsymbol{L}_{reg}(\boldsymbol{t_i},\boldsymbol{t_i^*}) \tag{4}$$

式中: i 为先验框序号; p_i 为此先验框被判定为目标的概率; $\boldsymbol{t_i}$ 为先验框尺寸参数向量; N_{obj} 为先验框数目,个; $\boldsymbol{L}_{obj}$ 为交叉熵损失函数; p_i^* 为先验框正负样本标志(1 为正样本, 0 为负样本); λ 为超参数, 是用于平衡二元分类任务和包围框回归任务权重的参数, 其值默认为 λ=5; N_{reg} 为目标框数目; $\boldsymbol{L}_{reg}$ 为 Smooth L1 损失函数; $\boldsymbol{t_i^*}$ 是此先验框对应样本值.

对背景框判定, 以鸟瞰图先验框和样本 2D 交并比(IoU)[16] 为判据, 具体如表 1 所示, 被判定为背景框的先验框不加入计算, 既不是目标框也不是背景框的不参与训练, 达到初步筛选的目的. 利用二维非极大值抑制算法进一步剔除冗余目标, 保留 IoU 阈值为 0.8 且最多不超过前 1 024 个的目标框, 以便提高召回率, 降低漏检.

表 1　样本 2D IoU 判定指标

Tab. 1　Sample 2D IoU judgment indicators

分类结果	车辆	行人	骑车人
目标框	IoU>0.5	IoU>0.5	IoU>0.5
背景框	IoU<0.3	IoU<0.45	IoU<0.45

1.2　目标检测网络

在得到粗略估计的候选框三维尺寸之后, 进行候选框尺寸的精细回归, 计算航向角与目标类别判断, 同时进行特征第二次融合.

对三维包围框尺寸估计时首先考虑编码方式.

主要有两种常用编码方式：利用六面体的 8 个顶点编码和轴对齐方式编码. 第一种方式能获得准确的尺寸估计，但是所需参数量较多；第二种利用中心点坐标和沿 3 个坐标轴的棱长编码（第一阶段使用），所需参数量较小，但是不能编码航向角信息. 本阶段采用新的编码方式对三维包围框尺寸编码，使用底面 4 点以及 2 个高度值（底面、顶面与地平面高度）的方式编码六面体，不仅可以获得准确的尺寸估计，而且所需参数量较小，编码方式如图 3 所示.

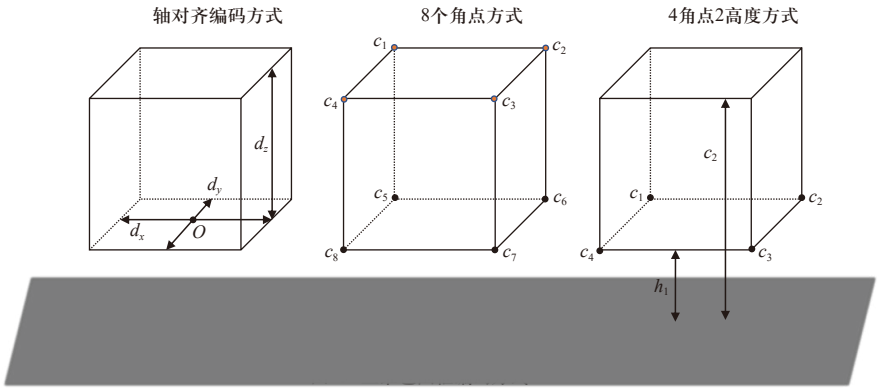


图 3 三维包围框编码方式

Fig. 3 Encoding method of 3D bounding box

回归后的目标共 10 个参数，相比于 8 个角点编码方式所需的 24 个参数大幅减少，回归后 10 个参数包括 8 个角点偏移量 $\Delta x,\Delta y$ 和 2 个高度偏移量 Δh（$\Delta x_1,\cdots,\Delta x_4,\Delta y_1,\cdots,\Delta y_4,\Delta h_1,\Delta h_2$）. 维护得到的角点，并约束 4 个角点构成一个矩形，选择各边中点，将对边中点连线，取较长边作为坐标轴基准，具体实现如图 4.

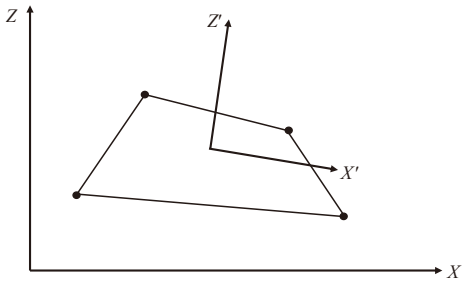


图 4 4 个角点确定方法

Fig. 4 Determination method of four corner points

AVOD 网络航向角编码方式如图 5 所示，计算方法基于一个二元向量隐式表达航向角，即 $(x_{or},y_{or})=(\cos\theta,\sin\theta)$，使 $[-\pi,\pi]$ 中的每一角度都有唯一单位向量相对应，保证航向角唯一性.

损失函数由三维包围框尺寸计算、航向角估计与目标类别三个任务损失函数构成. 使用原始的 256 通道特征图，将来自区域建议网络的候选框投影到特征图上获得候选特征，对投影后的特征图调整分辨率到 7×7 像素，并对元素取平均后融合. 融合后的特征通过三个每层 2 048 个节点的全连接层，分别

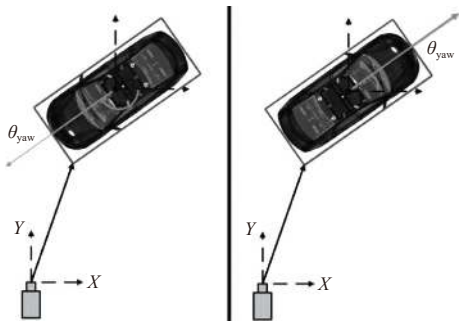


图 5 航向角编码方式示意图

Fig. 5 Schematic diagram of the heading angle coding method

输出三维包围框、航向角估计、目标类别，其中目标类别使用交叉熵代价函数来计算，其余两个使用 Smooth L1 损失函数计算. 最后对包围框筛选，利用 2D 非极大值抑制算法输出检测结果. 得到 AVOD 网络的损失函数计算式（5）：

$$\boldsymbol{L}(\{p_i\},\{\boldsymbol{t}_i\})=\frac{1}{N_{\text{cls}}}\sum\boldsymbol{L}_{\text{cls}}(p_i,p_i^*)+\frac{\lambda}{N_{\text{reg}}}\sum p_i^*\boldsymbol{L}_{\text{reg}}(\boldsymbol{t}_i,\boldsymbol{t}_i^*)+\frac{\lambda}{N_{\text{ang}}}\sum p_i^*\boldsymbol{L}_{\text{ang}}(\boldsymbol{t}_i,\boldsymbol{t}_i^*) \tag{5}$$

式中：$\boldsymbol{L}_{\text{cls}}$ 为交叉熵函数；$\boldsymbol{L}_{\text{reg}}$ 为 3D 包围框的 Smooth L1 损失函数；$\boldsymbol{L}_{\text{ang}}$ 为航向角估计的 Smooth L1 损失函数；N_{cls} 为先验框数目；N_{reg} 为目标框总数目. 根据鸟瞰图中 IoU 来判别候选框类型，对于车辆目标，鸟瞰图 IoU>0.65 时为正样本；对于行人和骑车人，IoU>0.55 为正样本，并参与到计算之中.

1.3　多视角标注信息联合损失函数

优化网络检测头部分损失函数计算: 将图像前视图与激光雷达点云鸟瞰特征图作为两个分支, 以各模态样本标注为基准监督学习, 计算各自损失函数, 针对性地优化特征提取网络, 防止图像特征提取网络退化, 框架如图 6 所示.

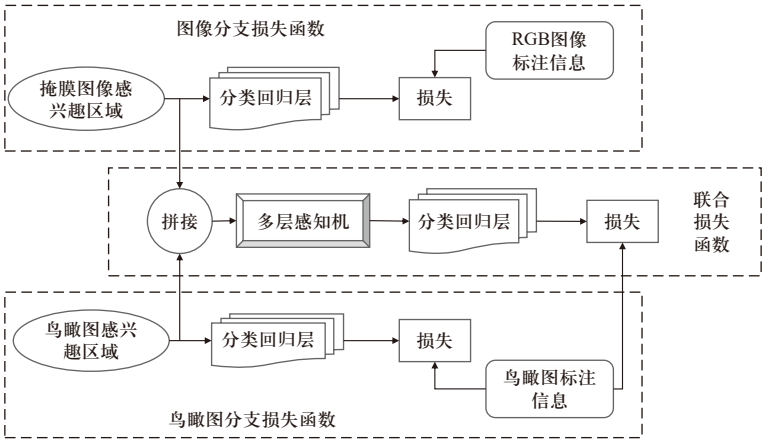


图 6　多视角标注信息联合损失函数

Fig. 6　Joint loss function of multi-view annotation information

对不同模态信息处理中加入全连接层, 首先进行包围框尺寸和目标类别的预判, 之后将预判结果与标注信息对比, 计算各模态损失函数.

$$\boldsymbol{L}_{\text{sub-cls}}=\frac{1}{N}\sum\boldsymbol{L}_{\text{cls}}\left(\boldsymbol{y}_i^{\text{img}},\widehat{\boldsymbol{y}}_i^{\text{img}}\right)+\frac{1}{N}\sum\boldsymbol{L}_{\text{cls}}(\boldsymbol{y}_i^{\text{bev}},\widehat{\boldsymbol{y}}_i^{\text{bev}})\tag{6}$$

$$\boldsymbol{L}_{\text{sub-reg}}=\frac{1}{N_p^{\text{img}}}\sum\boldsymbol{I}\left[\widehat{\boldsymbol{y}}_i^{\text{img}}>0\right]\boldsymbol{L}_{\text{reg}}(\boldsymbol{s}_i^{\text{img}},\widehat{\boldsymbol{s}}_i^{\text{img}})+\frac{1}{N_p^{\text{img}}}\sum\boldsymbol{I}\left[\widehat{\boldsymbol{y}}_i^{\text{bev}}>0\right]\boldsymbol{L}_{\text{reg}}(\boldsymbol{s}_i^{\text{bev}},\widehat{\boldsymbol{s}}_i^{\text{bev}})\tag{7}$$

式中: $\boldsymbol{L}_{\text{sub-cls}}$为分类模块损失函数; $\boldsymbol{L}_{\text{sub-reg}}$为包围框尺寸计算损失函数; N为目标框的总数量, N_p^{img}为前视图正样本数量, N_p^{bev}为鸟瞰图正样本数量, 单位(个); $\boldsymbol{I}$为选出正样本目标框的筛选函数, 为正值; $\boldsymbol{y}_i^{\text{img}}$、$\boldsymbol{y}_i^{\text{bev}}$分别为图像和鸟瞰图分支对第 i 目标框的分类估计值; $\widehat{\boldsymbol{y}}_i^{\text{img}}$和 $\widehat{\boldsymbol{y}}_i^{\text{bev}}$为图像和鸟瞰图的标注信息; $\boldsymbol{s}_i^{\text{img}}$、$\boldsymbol{s}_i^{\text{bev}}$为包围框尺寸偏移量和伸缩量; $\widehat{\boldsymbol{s}}_i^{\text{img}}$和 $\widehat{\boldsymbol{s}}_i^{\text{bev}}$为对应的标注信息.

对于正样本的判定, 基于包围框与标注信息框的交并比来划分. 在鸟瞰图中, 车辆类别的交并比大于 0.65 为正样本, 小于 0.55 为负样本, 行人与骑车人类别的交并比大于 0.45 为正样本, 小于 0.4 为负样本; 在前视图中, 车辆类别的交并比大于 0.7 为正样本, 小于 0.5 为负样本, 行人和骑车人类别交并比大于 0.6 为正样本, 小于 0.4 为负样本. 对于不属于正负样本的目标框来说, 不参与损失函数统计. 最终得到多视角标注信息网络的联合损失函数 AVOD-MLI (multi-view label information):

$$\begin{aligned}\boldsymbol{L}=&\frac{\lambda_{\text{cls}}}{N}\sum\boldsymbol{L}_{\text{cls}}(\boldsymbol{y}_i^{\text{fusion}},\widehat{\boldsymbol{y}}_i^{\text{bev}})+\\&\frac{\lambda_{\text{reg}}}{N_{\text{p}}^{\text{bev}}}\sum\boldsymbol{I}[\widehat{\boldsymbol{y}}_i>0]\,L_{\text{reg}}(s_i^{\text{fusion}},\widehat{s}_i^{\text{bev}})+\\&\frac{\lambda_{\text{ang}}}{N_p^{\text{bev}}}\sum\boldsymbol{I}\left[\widehat{\boldsymbol{y}}_i^{\text{bev}}>0\right]\boldsymbol{L}_{\text{ang}}(\boldsymbol{a}_i^{\text{fusion}},\widehat{\boldsymbol{a}}_i^{\text{bev}})+\\&\lambda_{\text{sub-cls}}\boldsymbol{L}_{\text{sub-cls}}+\lambda_{\text{sub-reg}}\boldsymbol{L}_{\text{sub-reg}}\end{aligned}\tag{8}$$

三维包围框的尺寸偏移和航向角的损失函数利用 Smooth L1 函数实现, 目标分类的损失函数利用交叉熵函数实现, λ作为超参数来权衡各任务损失函数权重.

2　双视角互投影融合方法的 AVOD-MPF 网络

2.1　AVOD-MPF 网络框架

AVOD 网络的数据融合发生在特征层, 通过拼接后按元素求平均的方式进行融合, 为了保证拼接时特征图分辨率一致, 经由池化层进行裁剪. 这种融合方式可能会使不同模态数据相互干扰, 从而削弱特征.

文中通过加入互投影池化层来改进网络的融合阶段, 改进后的网络可以优化不同模态数据特征融合, 充分利用了激光雷达点云的稀疏性, 将互投影池化层插入到 VGG 特征提取网络之后, 即特征金字塔的编码器之后, 解码器之前, 改进后的网络称为 AVOD-MPF(mutual projection fusion,MPF)网络, 局部网络结构如图 7.

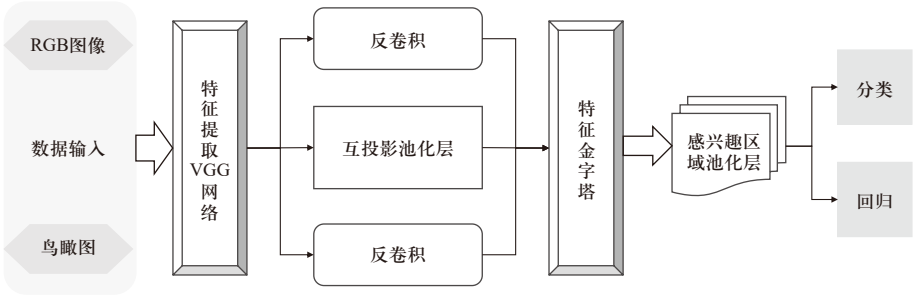


图 7 插入互投影池化层后局部网络结构

Fig. 7 Local network structure after inserting mutual projection pooling layer

2.2 互投影池化层

通过坐标互投影，将激光雷达点云变换到图像前视图，并将图像变换到激光雷达点云鸟瞰图，从而获得激光雷达点云在前视图的特征图以及图像在激光雷达点云鸟瞰图上的特征图，结构如图 8 所示. 通过相机与激光雷达的坐标转换矩阵$\boldsymbol{P}\in\mathbb{R}^{3\times4}$进行前视图与鸟瞰图之间的转换，如下式：

$$\boldsymbol{f}(\boldsymbol{x},\boldsymbol{y})=\sum_{u,v}\boldsymbol{k}_{x,y}(\boldsymbol{u},\boldsymbol{v})\boldsymbol{g}(\boldsymbol{u},\boldsymbol{v}) \tag{9}$$

$$\boldsymbol{k}_{x,y}=\{(\boldsymbol{u},\boldsymbol{v})|[\boldsymbol{u},\boldsymbol{v}]^{\mathrm{T}}=\omega^{-1}\boldsymbol{P}_{12}\boldsymbol{X},\omega\in\mathbb{R}^{+}\} \tag{10}$$

式中：(x,y)为鸟瞰图像素坐标；(u,v)为图像的像素坐标；$\boldsymbol{f}(\boldsymbol{x},\boldsymbol{y})$和$\boldsymbol{g}(\boldsymbol{u},\boldsymbol{v})$为两个特征图；$\boldsymbol{k}(\boldsymbol{u},\boldsymbol{v})$为运算核；$\boldsymbol{X}=[\boldsymbol{x}\ \ \boldsymbol{y}\ \ \boldsymbol{z}\ 1]^{\mathrm{T}}$，$\boldsymbol{P}_{12}$是$\boldsymbol{P}$前两行的子矩阵.

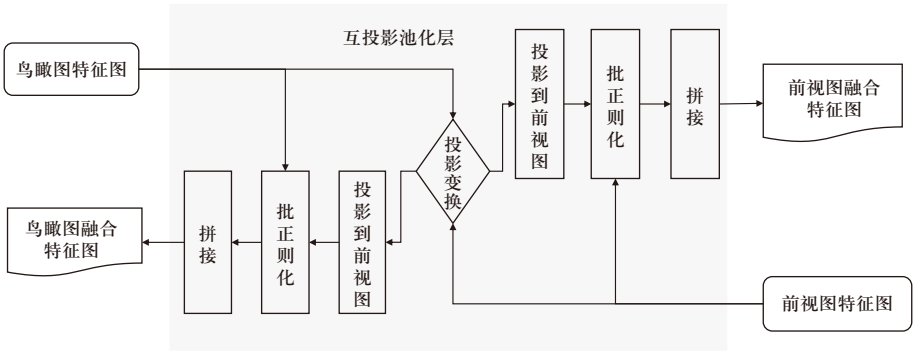


图 8 互投影池化层融合

Fig. 8 Mutual projection pooling layer fusion

通过上述转换会造成一个(u,v)对应多个(x,y)的状况，并且多个(x,y)点近似为直线$(\lambda x,\lambda y)$，影响运算，因此依据激光雷达点云的稀疏性进行稀疏化，进行非齐次转换. 假定前视图尺寸$\boldsymbol{L}_{\mathrm{f}}\times\boldsymbol{W}_{\mathrm{f}}$，鸟瞰图尺寸$\boldsymbol{H}_{\mathrm{b}}\times\boldsymbol{W}_{\mathrm{b}}$，激光点云记为$\{(\boldsymbol{x}_i,\boldsymbol{y}_i,\boldsymbol{z}_i),i=1,2,\cdots N\}$，则得到转换方程式为

$$\boldsymbol{k}_{x,y}(\boldsymbol{u},\boldsymbol{v})=\boldsymbol{\delta}_{(x,y)(x_i,y_i)}\boldsymbol{k}_{x,y}(\boldsymbol{u},\boldsymbol{v})\boldsymbol{\delta}_{(u,v)(u_i,v_i)},i,j=1,2,\cdots N \tag{11}$$

由多模态数据的对应性，可以将式(11)转化为

$$\begin{aligned}\boldsymbol{f}(\boldsymbol{x}_i,\boldsymbol{y}_i)=&\sum_{u,v}\boldsymbol{k}_{x_i,y_i}(\boldsymbol{u},\boldsymbol{v})\boldsymbol{\delta}_{(u,v)(u_i,v_i)}\boldsymbol{g}(\boldsymbol{u},\boldsymbol{v})=\\&\sum_{j}\boldsymbol{k}_{x_i,y_i}(\boldsymbol{u}_j,\boldsymbol{v}_j)\,\boldsymbol{g}(\boldsymbol{u}_j,\boldsymbol{v}_j)=\\&\sum_{j}\boldsymbol{k}_{x_i,y_i}(\boldsymbol{u}_j,\boldsymbol{v}_j)\boldsymbol{\delta}_{(x_i,y_i)(x_j,y_j)}\boldsymbol{g}(\boldsymbol{u}_j,\boldsymbol{v}_j)\end{aligned} \tag{12}$$

其中，

$$\boldsymbol{\delta}_{\mathrm{ab}}=\begin{cases}1,&a、b对应同一点\\0,&其他\end{cases} \tag{13}$$

多传感器数据投影是双向进行的，可以在不同视角下形成特征图. 在拼接前对每一特征图进行归一化处理，利用批正则化层来实现；将前视图特征图与稀疏矩阵相乘并与鸟瞰图数据的特征图相拼接融合，类似的鸟瞰图的特征图以同样方式与前视图特征图相拼接融合. 其中稀疏矩阵$\boldsymbol{X}$尺寸为$\boldsymbol{L}_{\mathrm{f}}\boldsymbol{W}_{\mathrm{f}}\times\boldsymbol{H}_{\mathrm{b}}\boldsymbol{W}_{\mathrm{b}}$，尺寸为$\boldsymbol{H}_{\mathrm{b}}\times\boldsymbol{W}_{\mathrm{b}}\times\boldsymbol{A}$的特征图转化后为尺寸$\boldsymbol{H}_{\mathrm{b}}\boldsymbol{W}_{\mathrm{b}}\times\boldsymbol{A}$的矩阵$\boldsymbol{F}$. 最终得到特征图$\boldsymbol{T}=\boldsymbol{MF}$，其尺寸为$\boldsymbol{L}_{\mathrm{f}}\boldsymbol{W}_{\mathrm{f}}\times\boldsymbol{A}$.

3 验证与分析

3.1 三维目标检测评价指标

KITTI 数据集[17]本身根据各种状况将目标进行

划分,三种难度级别为:简单(最小包围框高度≥40像素,目标完全可见,截断≤15%)、中等(最小包围框高度≥25像素,目标部分可见,截断≤30%)、困难(困难:最小包围框高度≥25像素,目标难以看见,截断≤50%),划分依据主要是目标大小、遮挡以及截断情况.

文中网络主要针对车辆、行人、骑车人进行目标检测,并在验证集上统计标注样本和目标检测结果,利用三维平均精度 AP_{3D} 来评价目标检测网络在三维尺度的检测精度.

3.2　训练策略与设备

对于目标检测任务,KITTI 数据集拥有大量的图像和激光雷达点云数据用于训练,针对不同尺度的目标训练了两种模型,分别对应车辆、行人以及骑车人,为了确保网络改进结果的合理性与有效性,分别对原 AVOD 网络、AVOD-MLI 网络、AVOD-MPF 网络分别进行训练并进行结果对比.

文中所用的训练机配置为:32 GB 内存,11 G 显存,Nvidia 1080Ti 显卡,IntelCore i7-8700K @3.70 GHz × 12 的 CPU,在 Ubuntu 16.04 操作系统下进行,深度学习框架为 Tensorflow.

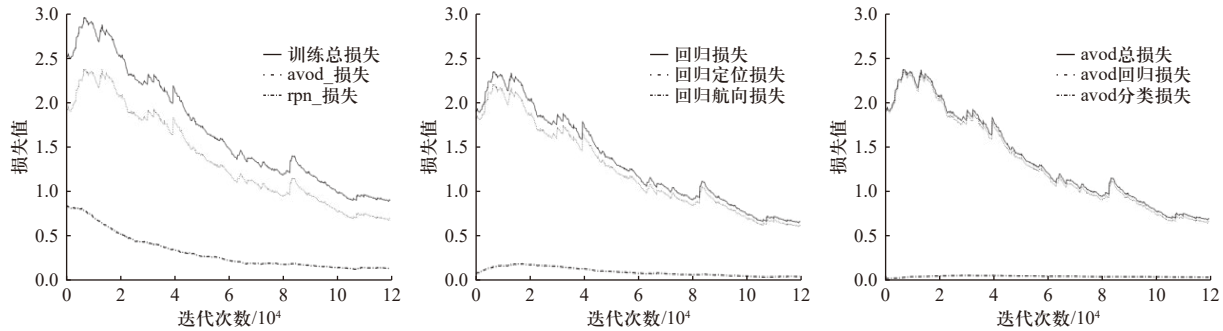
3.3　训练过程与结果验证

3.3.1　AVOD 网络训练及结果

使用 ADAM 优化器对模型参数进行优化,设定初始学习率为 0.000 1,指数衰减,共 120 k 次迭代训练,每 100 k 次训练进行一次衰减,衰减系数定为 0.1.全连接层引入 dropout 方法,并利用批正则化方法.区域建议网络中设定建议框样本为 512 个,第 2 阶段的检测网络 1 024 个样本,正负样本各一半,并将此两个网络进行联合训练.

网络训练的总损失值由两部分:检测网络损失(AVOD 损失)以及区域建议网络损失(RPN 损失)构成,如图 9 (a) 所示.其中 AVOD 损失包含 AVOD 回归损失与 AVOD 分类损失,如图 9 (b) 所示,回归损失为主要影响要素,占据 AVOD 损失极大比例,且趋势与之相似.对 AVOD 回归损失进行分析,如图 9 (c) 所示,包含回归定位损失与回归航向角损失.模型在训练中损失值随迭代次数增加而呈下降和收敛之势,最后训练损失到达 0.615 2.

选取 F-pointNet 网络进行对比,结果如表 2 所示.



(a) 网络总损失　(b) AVOD总损失　(c) AVOD回归损失

图 9　AVOD 网络训练过程中损失函数值的变化

Fig. 9　Changes in loss function values during AVOD network training

表 2　AVOD 网络在验证集上的 AP_{3D}

Tab. 2　AP_{3D} of AVOD network on validation set

网络	车辆			行人			骑车人		
	困难	中等	简单	困难	中等	简单	困难	中等	简单
AVOD	68.96	75.13	85.27	40.07	43.23	51.21	46.04	52.98	65.27
F-pointNet	60.59	69.79	82.09	38.08	42.15	50.46	49.01	56.08	71.13

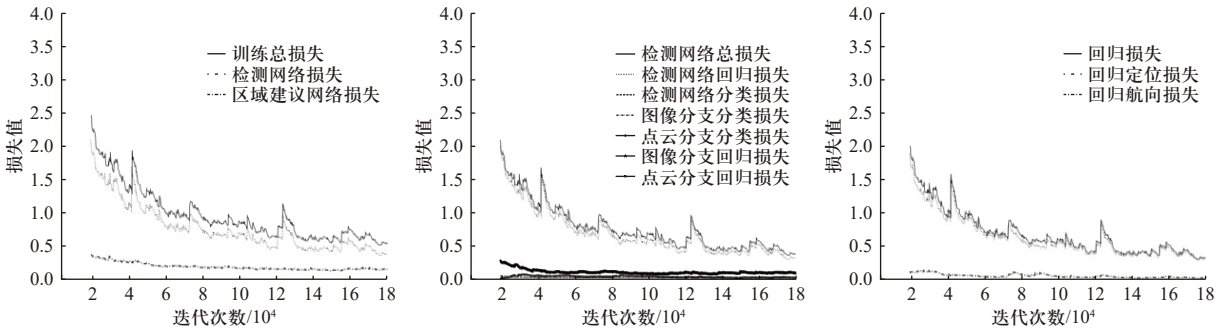
表中可以看出 AVOD 网络具有更好的车辆目标检测精度,对于困难和中等类别的车辆检测精度分别提高了 8.37%、5.34%,处理有遮挡和截断目标的效果较好,但是对于小尺度目标的检测精度较低.

3.3.2　AVOD-MLI 网络训练及结果

使用 ADAM 优化器,初始学习率 0.000 1,指数衰减,每 100 k 次迭代进行一次衰减,衰减因子 0.1.使用最小批尺寸为 1 的 Xavier 对鸟瞰图特征提取网络初

始化，图像数据特征提取网络加入预训练的 ImageNet 权重. 区域建议网络仍设定 512 个建议框样本，第二阶段的检测网络 1 024 个样本，正负样本各一半.

AVOD-MLI 网络训练时损失值如图 10 所示.



(a) 网络总损失 (b) AVOD-MLI检测网络总损失 (c) AVOD-MLI回归损失

图 10 AVOD-MLI 网络训练损失值变化曲线

Fig. 10 AVOD-MLI network training loss value change curve

AVOD-MLI 网络的 $AP_{3D,}$ 与 AVOD 网络对比如 表 3 所示.

表 3 AVOD-MLI 网络在 KITTI 数据集上的 AP_{3D}

Tab. 3 AP_{3D} of AVOD-MLI network on KITTI dataset

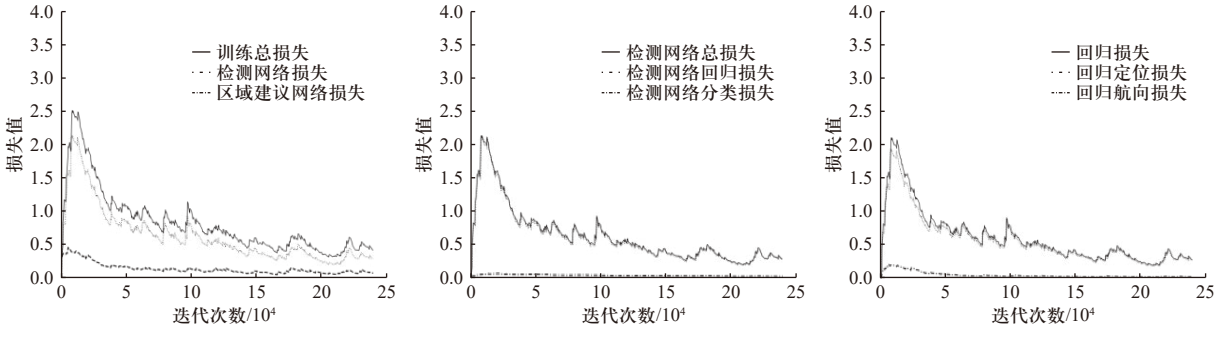
网络	车辆			行人			骑车人		
	困难	中等	简单	困难	中等	简单	困难	中等	简单
AVOD-MLI	69.42	74.94	84.41	41.25	45.02	53.96	47.21	54.16	66.72
AVOD	68.96	75.13	85.27	40.07	43.23	51.21	46.04	52.98	65.27

从表 3 可以看出 AVOD-MLI 网络对于车辆目标效果不明显，可能是由于对于车辆目标而言，较大的尺寸差异造成激光雷达点云特征图的分支占据了较大比例，图像分支未能带来显著提升. 对于小尺度目标来说，激光雷达点云特征反而被削弱，图像特征能够带来更多的纹理信息，因此对于行人目标和骑车人目标而言，提升更为明显，对于行人目标，不同难度目标分别提高了 1.18%, 1.79%, 2.75%.

3.3.3 AVOD-MPF 网络训练及结果

引入 ADAM 优化器，初始学习率 0.000 1，指数衰减，每 30k 次迭代进行一次衰减，衰减因子 0.8.使用最小批尺寸为 1 的 Xavier 对鸟瞰图特征提取网络初始化，图像数据特征提取网络加入预训练的 ImageNet 权重. 区域建议网络仍设定 512 个建议框样本，第 2 阶段的检测网络 1 024 个样本，正负样本各一半.

AVOD-MPF 网络训练时损失值如图 11 所示. 随着迭代次数增加，网络总损失逐渐收敛，终值为 0.269 5，如图 10(a) 所示，AVOD-MPF 检测网络变化值与 AVOD-MPF 回归损失分别如图 11(b)、11(c) 所示，学习率如图 12 所示.



(a) 网络总损失 (b) AVOD-MPF检测网络总损失 (c) AVOD-MPF回归损失

图 11 AVOD-MPF 网络训练损失值变化曲线

Fig. 11 AVOD-MPF network training loss value change curve

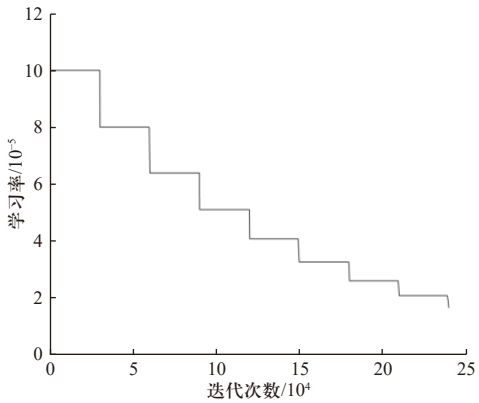


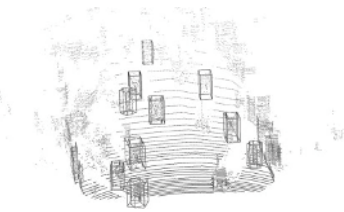
图 12　AVOD-MPF 网络训练学习率

Fig. 12　AVOD-MPF network training learning rate

行人和车辆检测如图 13 和图 14 所示, 其中 13(a) 和 14(a) 为最终 3D 目标检测结果; 图 13(b) 和 14(b) 为第 1 阶段网络处理结果, 实线框为建议框, 虚线框为检测框; 图 13(c) 和 14(c) 为整体网络回归结果, 虚线框为标注框, 实线框为检测框.

(a) 图像检测框

(b) 建议框和检测框

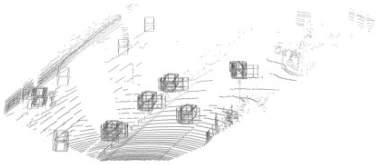


(c) 检测框和标注框

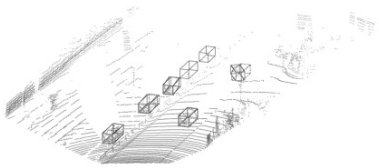
图 13　行人检测示例

Fig. 13　Example of pedestrian detection

(a) 图像检测框



(b) 建议框和检测框



(c) 检测框和标注框

图 14　车辆检测示例

Fig. 14　Vehicle detection example

AVOD-MPF 网络的 AP_{3D} 与 AVOD 网络对比如表 4 所示.

表 3 和表 4 数据显示加入互投影池化层的 AVOD-MPF 网络保留了 AVOD 网络本身对于车辆目标检测的优势, 相比于 F-pointNet 网络, 对遮挡严重的车辆目标检测精度提高了 8.62%. 同时提高了 AVOD 网络对小尺度目标的检测精度, 对于中等难度目标来说, AVOD-MPF 网络将行人检测精度提高了 2.03%, 骑车人检测精度提高了 2.34%, 说明加入的互投影池化层能够提升网络性能, 改善了原 AVOD 网络小尺度目标检测精度不高的问题.

表 4　AVOD-MPF 网络在 KITTI 数据集上的 AP_{3D}

Tab. 4　AP_{3D} of AVOD-MPF network on KITTI dataset

网络	车辆			行人			骑车人		
	困难	中等	简单	困难	中等	简单	困难	中等	简单
AVOD-MPF	69.21	74.79	86.37	42.10	44.64	52.57	47.72	55.32	67.85
AVOD-MLI	69.42	74.94	84.41	41.25	45.02	53.96	47.21	54.16	66.72
AVOD	68.96	75.13	85.27	40.07	43.23	51.21	46.04	52.98	65.27
F-pointNet	60.59	69.79	82.09	38.08	42.15	50.46	49.01	56.08	71.13

4 结 论

提出一种基于视觉与激光雷达的多视角互投影融合的三维目标检测方法，改进对车辆检测精度较高的 AVOD 网络，通过互投影的方式加强不同模态信息数据关联并进行特征级融合. 相比于其他算法以及原网络来说，文中使用的 AVOD-MPF 网络方法具有明显优势，实验数据和结果表明，本方法不仅能够实现三维目标检测时特征级和决策级融合，而且在保留 AVOD 网络对车辆目标检测优势的同时，也提升了对行人和骑车人等小尺度目标的检测精度，对于有遮挡的目标复杂场景也有较好的适应性，为小尺度目标检测提供了一种新的思路.

参考文献：

[1] 刘语佳. 基于图像和点云的目标检测和跟踪方法研究[D]. 北京: 北京理工大学, 2021.
LIU Yujia. Object detection and tracking based on fusion of image and point cloud[D]. Beijing: Beijing Institute of Technology, 2021. (in Chinese)

[2] 于加其, 杨树兴, 朱伯立. 弹载激光成像雷达距离像的目标提取技术[J]. 北京理工大学学报, 2016, 36(12): 1279 − 1282.
YU Jiaqi, YANG Shuxing, ZHU Boli. Target extraction base on range image from missile-borne imaging LADAR[J]. Transactions of Beijing Institute of Technology, 2016, 36(12): 1279 − 1282. (in Chinese)

[3] 万鹏. 基于F-PointNet的3D点云数据目标检测[J]. 山东大学学报(工学版), 2019, 49(5): 98 − 104.
WAN Peng. Object detection of 3D point clouds based on F-PointNet[J]. Journal of Shandong University(Engineering Science Edition), 2019, 49(5): 98 − 104. (in Chinese)

[4] ASVADI A, GARROTE L, PREMEBIDA C, et al. Multimodal vehicle detection: fusing 3D-LIDAR and color camera data[J]. Pattern Recognition Letters, 2018, 115: 20 − 29.

[5] 姚钺, 任明武. 基于PointNet++改进的点云特征提取与分类网络架构[J]. 计算机与数字工程, 2021, 49(10): 2052 − 2056,2112.
YAO Yue, REN Mingwu. Improved point cloud feature extraction and classification network architecture based on PointNet++[J]. Computer & Digital Engineering, 2021, 49(10): 2052 − 2056,2112. (in Chinese)

[6] VORA S, LANG A H, HELOU B, et al. Pointpainting: sequential fusion for 3d object detection[C]//Proceedings of Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE/CVF, 2020: 4604 − 4612.

[7] 杨柳, 刘启亮, 袁浩涛. 城市激光点云语义分割典型方法对比研究[J]. 地理与地理信息科学, 2021, 37(1): 17 − 25.
YANG Liu, LIU Qiliang, YUAN Haotao. A comparative study on typical methods for semantic segmentation of laser point clouds in urban areas[J]. Geography and Geo-Information Science, 2021, 37(1): 17 − 25. (in Chinese)

[8] YAN Y, MAO Y, LI B. Second: sparsely embedded convolutional detection[J]. Sensors, 2018, 18(10): 3337.

[9] GUJJAR H S. A comparative study of VoxelNet and PointNet for 3D object detection in car by using KITTI benchmark[J]. International Journal of Information Communication Technologies and Human Development (IJICTHD), 2018, 10(3): 28 − 38.

[10] STANISZ J, LIS K, GORGON M. Implementation of the PointPillars network for 3D object detection in reprogrammable heterogeneous devices using FINN[J]. Journal of Signal Processing Systems, 2021, 12: 1 − 16.

[11] CHEN X, KUNDU K, ZHU Y, et al. 3D object proposals for accurate object class detection[C]//Proceedings of the 28th International Conference on Neural Information Processing Systems-Volume 1. Istanbul, Turkey. Springer, 2015: 424 − 432.

[12] SINDAGI V A, ZHOU Y, TUZEL O. Mvx-net: multimodal voxelnet for 3d object detection[C]//Proceedings of the 2019 International Conference on Robotics and Automation (ICRA). Montreal, Canada. IEEE, 2019: 7276 − 7282.

[13] SONG S, XIAO J. Sliding shapes for 3D object detection in depth images[C]//Proceedings of the European conference on computer vision. Zurich, Switzerland. Springer, Cham, 2014: 634 − 651.

[14] 陈思颖, 王晓鲁, 张寅超等. 城区机载激光雷达点云数据与航空影像的多尺度配准方法[J]. 北京理工大学学报, 2016, 36(2): 186 − 190.
CHEN Siying, WANG Xiaolu, ZHANG Yinchao, et al. Multi-Scale registration of LiDAR data and aerial image[J]. Transactions of Beijing Institute of Technology, 2016, 36(2): 186 − 190. (in Chinese)

[15] 汪常建, 丁勇, 卢盼成. 融合改进FPN与关联网络的Faster R-CNN目标检测[J]. 计算机工程, 2022, 48(2): 173 − 179.
WANG Changjian, DING Yong, LU Pancheng. Object detection using faster R-CNN combining improved FPN and relation network[J]. Computer Engineering, 2022, 48(2): 173 − 179. (in Chinese)

[16] YAN J, WANG H, YAN M, et al. IoU-adaptive deformable R-CNN: make full use of IoU for multi-Class object detection in remote sensing imagery[J]. Remote Sensing, 2019, 11(3): 286.

[17] GEIGER A, LENZ P, STILLER C, et al. Vision meets robotics: The KITTI dataset[J]. The International Journal of Robotics Research, 2013, 32(11): 1231 − 1237.

（责任编辑：孙竹凤）