doi:10.11835/j.issn.1000.582X.2024.04.012

面向自动驾驶的多模态信息融合动态目标识别

张明容[1],喻　皓[2],吕　辉[3],姜立标[3],李利平[3],卢　磊[4]
(1. 广东轻工职业技术学院 汽车技术学院,广州 510000; 2. 广汽埃安新能源汽车股份有限公司研发中心,广州 511400; 3. 华南理工大学 机械与汽车工程学院,广州 510641; 4. 广州城市理工学院 工程研究院,广州 510800)

摘要:研究提出一种面向自动驾驶的多模态信息融合的目标识别方法,旨在解决自动驾驶环境下车辆和行人检测问题。该方法首先对ResNet50网络进行改进,引入基于空间注意力机制和混合空洞卷积,通过选择核卷积替换部分卷积层,使网络能够根据特征尺寸动态调整感受野的大小;然后,卷积层中使用锯齿状混合空洞卷积,捕获多尺度上下文信息,提高网络特征提取能力。改用GIoU损失函数替代YOLOv3中的定位损失函数,GIoU损失函数在实际应用中具有较好操作性;最后,提出了基于数据融合的人车目标分类识别算法,有效提高目标检测的准确率。实验结果表明,该方法与OFTNet、VoxelNet和FasterRCNN网络相比,在mAP指标白天提升幅度最高可达0.05,晚上可达0.09,收敛效果好。

关键词:自动驾驶;ResNet50;YOLOv3;数据融合;注意力机制;损失函数
中图分类号:T391　　**文献标志码:**A　　**文章编号:**1000-582X(2024)04-139-18

Multimodal information fusion dynamic target recognition for autonomous driving

ZHANG Mingrong[1], *YU Hao*[2], *LYU Hui*[3], *JIANG Libiao*[3], *LI Liping*[3], *LU Lei*[4]
(1. School of Automotive Technology, Guangdong Industry Polytechnic, Guangzhou 510000, P. R. China; 2. GAC AION New Energy Automobile Co., Ltd., Guangzhou 511400, P. R. China; 3. School of Mechanical & Automotive Engineering, South China University of Technology, Guangzhou 510641, P. R. China; 4. Engineering Research Institute, Guangzhou City University of Technology, Guangzhou 510800, P. R. China)

Abstract: A multi-modal information fusion based object recognition method for autonomous driving is proposed to address the vehicle and pedestrian detection challenge in autonomous driving environments. The method first improves ResNet50 network based on spatial attention mechanism and hybrid null convolution. The standard convolution is replaced by selective kernel convolution, which allows the network to dynamically adjust the size of the perceptual field according to the feature size. Then, the sawtooth hybrid null convolution is used to enable the network to capture multi-scale contextual information and improve the network feature extraction capability.

收稿日期:2023-05-12
基金项目:国家自然科学基金资助项目(51975217)。
Supported by National Natural Science Foundation of China(51975217).
作者简介:张明容(1983—),女,博士,副教授,主要从事智能网联汽车方向研究,(E-mail)153155269@qq.com。
通信作者:喻皓,男,高级工程师,(E-mail)yuhao@gacne.com.cn。

The localization loss function in YOLOv3 is replaced with the GIoU loss function, which has better operability in practical applications. Finally, human-vehicle target classification and recognition algorithm based on two kinds of data fusion is proposed, which can improve the accuracy of the target detection. Experimental results show that compared with OFTNet, VoxelNet and FASTERRCNN, the mAP index can be improved by 0.05 during daytime and 0.09 in the evening, and the convergence effect is good.

Keywords: autonomous driving; ResNet50; YOLOv3; data fusion; attention mechanism; loss function

随着互联网企业、造车新势力以及传统车企纷纷投入自动驾驶市场,自动驾驶领域呈现火热势态。自动驾驶汽车,又称无人驾驶汽车、电脑驾驶汽车或轮式移动机器人,其系统主要由感知、决策、控制3部分组成[1]。

自动驾驶中用于环境感知的数据主要来源于图像传感器和激光雷达,图像传感器作为一种被动式传感器,成像质量受外界光照影响较大,无法在过曝、黑夜以及恶劣天气如雾霾、暴雪等极端光照条件下完成感知任务[2]。激光雷达(light detection and ranging,LiDAR)作为一种主动式光学传感器,对光照具有较好鲁棒性,具有精度高、范围大、抗有源干扰能力强的特性。但受限于技术条件,激光雷达获取的数据存在稀疏无序、难以直接利用的特点,且缺乏颜色和纹理信息,单靠激光雷达数据很难完成如车辆识别、行人检测等高级感知任务。由于驾驶环境复杂多变,单一传感器存在自身缺陷,只依赖于LiDAR或图像传感器难以保证检测的稳定性和可靠性,因此,笔者提出基于多模态信息融合的交通态势感知平台主要包含以下模块,如图1所示。

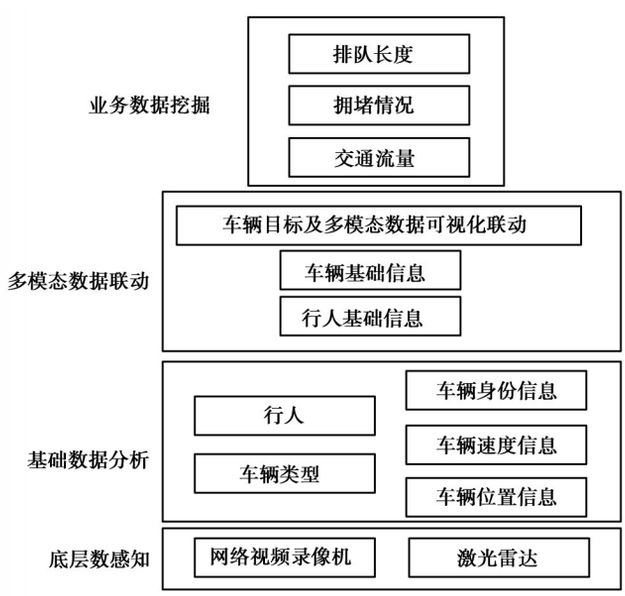


图1 交通态势感知平台总体架构

Fig. 1 Overall architecture of traffic situational awareness platform

结合激光雷达点云数据对环境的精准定位和RGB图像丰富的语义信息,可将这类方法分为早融合(Early Fusion)、深度融合(Deep Fusion)、晚融合(Late Fusion)三类[3]。

Early Fusion以Point Painting为代表,这是一种由Vora等人[4]提出用图像语义分割的结果来给点云“着色”的方法。在Late Fusion中,多种模态一般都分别拥有各自骨干网进行特征提取,随后利用共享候选框进行感兴趣区域池化(ROI pooling)[5]。Chen等人提出的MV3D[3]则是这类方法的典型。MV3D是一种多视角的3D目标检测网络,该方法使用BEV点云、FV点云以及FV图像作为输入。由于BEV图中遮挡情况最少,

所以在BEV中进行特征提取并送入RPN网络，将ROI向另外两图进行映射，得到3组ROI使用Deep Fusion的方式进行特征融合。Ku，Mozifian等人[6]则在MV3D基础上进一步提出了AVOD。区别于MV3D使用ROI pooling来处理多种视角特征图尺寸的一致性问题，AVOD则直接使用裁剪与尺寸调整的方式。

1　相关内容

近年来，国外激光雷达与视觉的目标检测研究取得了显著进展。Botha等人[7]（2017年）提出一种先进的数据融合方法，通过整合雷达和立体视觉数据，成功实现对运动目标的高效检测和跟踪。这项研究充分利用雷达和视觉传感器的互补性，有效提高目标检测的准确性和鲁棒性。Li等人[8]（2020年）的研究集中于激光雷达点云在自动驾驶中的应用。通过深度学习技术，研究人员能更精准分析和理解激光雷达点云数据，为自动驾驶系统提供更可靠的感知能力。2017年研究者们基于2D激光扫描仪和机器视觉的信息融合，致力于葡萄藤sucker的识别与定位，为农业领域的实际问题提供了解决方案[9]。Barrientos等人[10]（2013年）提出一种移动机器人上的人体检测方法，通过激光和视觉信息融合，实现对人体的有效探测。这种技术在机器人应用中具有广泛潜在用途，特别是在导航和安全领域。也有学者使用了3D和2D视觉信息融合的方案，实现准确定位和跟踪[11]，这一创新性方法为高精度计算机视觉应用提供了可靠技术支持。

近几年，基于深度卷积神经网络的目标识别技术得到飞速发展，检测性能也得到极大提高。Guda等人[12]提出了一阶段目标检测算法的开篇之作YOLOv1，YOLO系列的目标检测算法受到高度关注，后出现了YOLOv2、YOLOv3的目标检测算法，通过在原始网络的基础上不断找到创新技术并解决上一个版本遗留下来的问题，YOLO系列的目标检测算法不只是在理论研究上火热，更被应用到无数工业检测任务中，取得令人满意效果。

2　实验模型

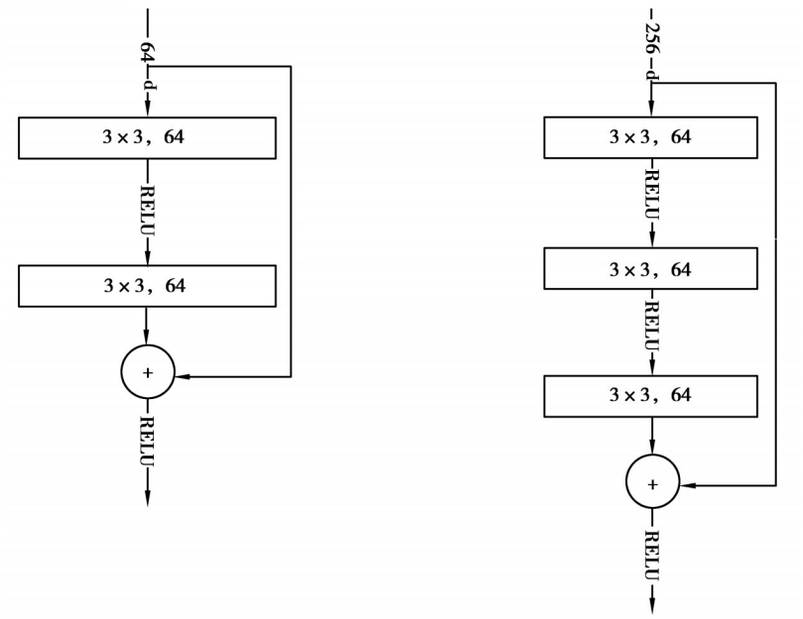
2.1　基于注意力机制改进的ResNet50道路目标特征提取

网络的性能受网络深度、宽度和卷积核尺寸等因素的影响，扩展网络宽度和卷积核尺寸对硬件设备要求高，而通过堆叠卷积层来增加网络深度，训练时会产生梯度消失现象，导致网络难以训练，性能出现退化。在极端情况下，增加的网络层即使学习不到有用信息，也可以将浅层网络学习的特征传递给全连接层，保证训练时网络性能不退化，这样的新层具有恒等映射（Identitymapping）功能。何凯明等[13]根据此思想提出了基于残差模块的ResNet网络。ResNet网络在实验室中可训练的深度已超过1 000层，但常用深度共有18/34/50/101/152五种。何凯明等人在实现ResNet网络时，考虑到计算成本，设计了block和bottleneck两种残差模块，分别对应ResNet18/34和ResNet50/101/152。ResNet50对应bottleneck残差模块，bottleneck使用1×1+3×3+1×1卷积结构。先利用第一个尺寸为1×1的卷积进行降维，然后在第二个尺寸为1×1的卷积中还原维度，达到计算精度不变，且能够降低计算量的目的。bottleneck残差模块的参数量是block残差模块的1/16.94。

研究使用ResNet50进行街道场景特征提取，对ResNet50网络进行改进设计，改进部分集中在网络的特征提取部分。ResNet50网络由conv1、conv2_x、conv3_x、conv4_x、conv5_x和一个全连接层组成，下图展示了ResNet50的网络结构，其中conv1是卷积核大小为7×7的标准卷积，conv2_x、conv3_x、conv4_x和conv5_x部分由残差模块堆叠而成，数量分别为3、4、6、3，每一部分的残差模块都可以根据需要更改参数，模型的模块化性能优越。

ResNet50是通过增加深度来提高模型的特征提取能力，它由bottleneck残差模块堆叠而成。bottleneck残差模块是通过三层标准卷积来实现对输入数据的特征提取，其第一层与第三层卷积，卷积核大小均为1×1，在特征提取过程中起辅助作用。第一层1×1卷积对输入数据进行降维处理，第二层1×1卷积还原数据维度，使得bottleneck残差模块与block残差模块相比，运算过程中既保证了计算精度，也降低了参数量。ResNet50网络通过不同的步长设计，随着网络深度增加，卷积的感受野越来越大，提取的特征越来越具有全局性，在这个过程中，使用标准卷积的残差模块对图片中每一部分关注度相同，固定的感受野大小只能学习

到相应尺寸的图片特征。



(a) Block残差模块　　(b) Bottleneck残差模块

图2　block和bottleneck结构

Fig. 2　block and bottleneck structure

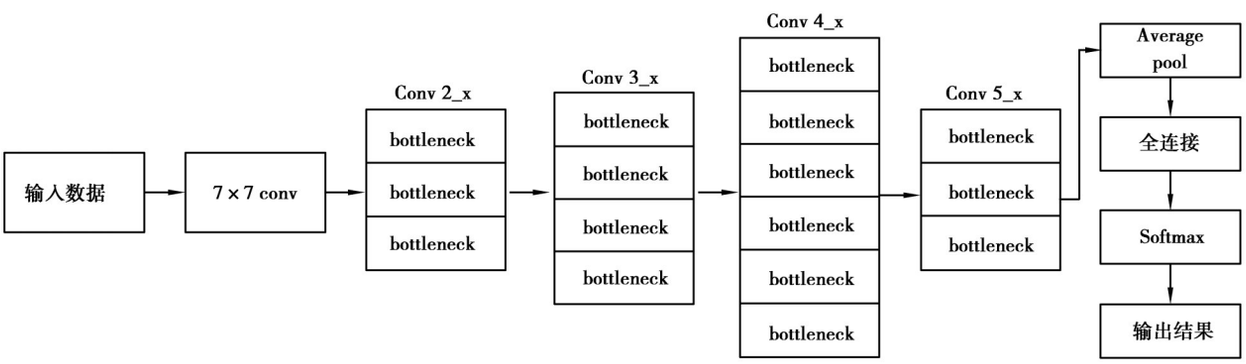


图3　ResNet50网络结构图

Fig. 3　ResNet50 network structure diagram

2.1.1　注意力机制

注意力机制基本思想是关注重点信息、抑制无用信息，增强网络提取特征的效率和准确度。注意力机制根据作用域不同，可分为通道注意力机制、空间注意力机制和混合域注意力机制。选择核卷积由分裂、融合、选择3步组成。

1)分裂操作如图4所示，对于给定的输入特征映射$X \in \mathbb{R}^{H' * W' * C'}$，通过卷积核大小为3×3，扩张率分别为1、2和3的3个分组卷积转换，得到3个感受野大小不同的特征图：$U_1 \in \mathbb{R}^{H * W * C}$，$U_2 \in \mathbb{R}^{H * W * C}$和$U_3 \in \mathbb{R}^{H * W * C}$。3条支流均由分组卷积、批量归一化和ReLU激活函数共同组成。

2)融合操作如图5所示，首先将3个特征图相加

$$U = U_1 + U_2 + U_3。 \tag{1}$$

然后使用全局平均池化层嵌入全局信息，得到通道尺度上具有全局信息的向量$\boldsymbol{s}$。

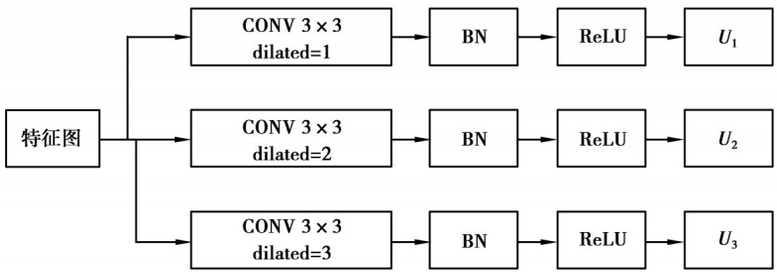


图 4　分裂操作示意图

Fig. 4　Schematic diagram of splitting operation

$$\boldsymbol{s}=F_{\mathrm{gp}}(U_{\mathrm{c}})=\frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W}U_{\mathrm{c}}(i,j)。\tag{2}$$

最后再经过一层全连接层,生成紧凑特征z。

$$z=F_{fc}(\boldsymbol{s})=\delta_{\mathrm{relu}}(B(W\boldsymbol{s})),\tag{3}$$

其中:δ表示ReLU激活函数,B表示批量归一化、$W\in\mathbb{R}^{d\times c}$。下式中$r$和$L$用来控制输出向量的维度,一般情况下,$L=32$。

$$d=\max\left(\frac{c}{r},L\right)。\tag{4}$$

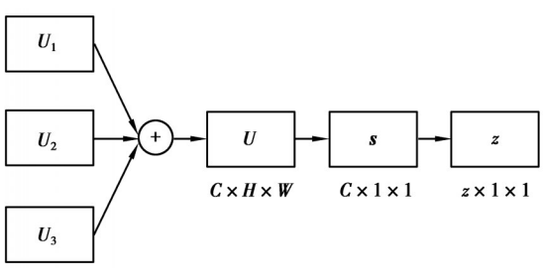


图 5　融合操作示意图

Fig. 5　Schematic diagram of fusion operation

3)选择操作如图6所示,基于softmax方法,利用紧凑特征z指导注意力机制动态选择不同感受野大小的信息。每条支流的权重向量计算方法如下

$$\begin{cases}\boldsymbol{a}_n=\dfrac{e^{A_{n^z}}}{e^{A_{n^z}}+e^{B_{n^z}}+e^{C_{n^z}}},\\ \boldsymbol{b}_n=\dfrac{e^{B_{n^z}}}{e^{A_{n^z}}+e^{B_{n^z}}+e^{C_{n^z}}},\\ \boldsymbol{c}_n=\dfrac{e^{C_{n^z}}}{e^{A_{n^z}}+e^{B_{n^z}}+e^{C_{n^z}}},\end{cases}\tag{5}$$

其中:$\boldsymbol{a}$、$\boldsymbol{b}$和$\boldsymbol{c}$分别表示特征图U_1、U_2和U_3的软注意力机制向量。而$\boldsymbol{a}_n$表示$\boldsymbol{a}$的第n个序列值,$\boldsymbol{b}_n$表示$\boldsymbol{b}$的第n个序列值,$\boldsymbol{c}_n$表示$\boldsymbol{c}$的第n个序列值,$A,B,C\in\mathbb{R}^{c\times d}$,$A_n$表示$A$的第$n$行,$B_n$表示$B$的第$n$行,$C_n$表示$C$的第$n$行。通过将权重向量$\boldsymbol{a}$、$\boldsymbol{b}$和$\boldsymbol{c}$分别和特征图$U_1$、$U_2$和$U_3$进行加权求和,获得输出向量$\boldsymbol{V}=\{V_1,V_2...V_n\}$,$\boldsymbol{V}_n\in\mathbb{R}^{H\times W}$。

$$\boldsymbol{V}_n=\boldsymbol{a}_n\times U_1+\boldsymbol{b}_n\times U_2+\boldsymbol{c}_n\times U_3。\tag{6}$$

在ResNet50网络中引入空间注意力机制,即将选择核卷积替换bottleneck残差模块中的标准卷积,图7展示了使用选择核卷积的bottleneck残差模块。在ResNet50网络中使用选择核卷积,可筛选特征信息,提高数据利用效率,且在选择核卷积的融合操作部分,卷积核尺寸不同的3组分组卷积既可使网络提取的特征更多样,增加ResNet50网络的宽度。选择核卷积只对于卷积核>1的标准卷积改造有效,选择使用选择核卷积

替换 bottleneck 残差模块中的第二层卷积，其卷积核大小为 3×3。

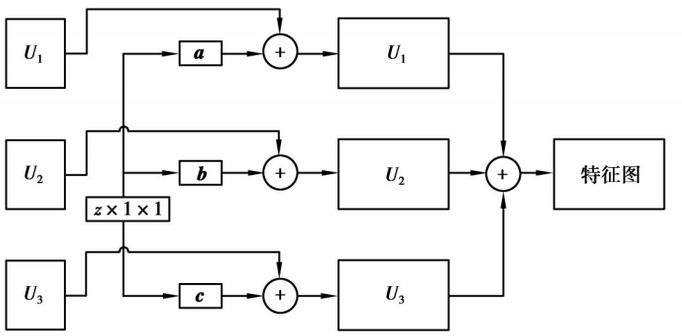


图 6 选择操作示意图

Fig. 6 Schematic diagram of the selection operation

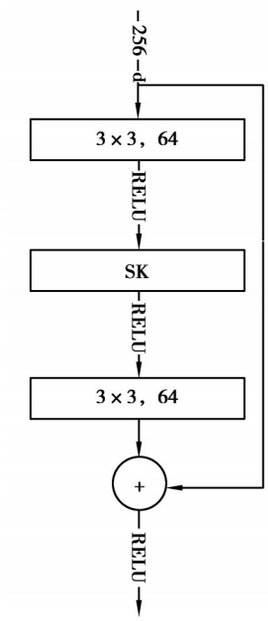


图 7 引入空间注意力机制的残差模块

Fig. 7 Residual module with the introduction of spatial attention mechanism

在 ResNet50 网络的 conv2_x、conv3_x 部分，使用选择核卷积替换 bottleneck 残差模块中的 3×3 标准卷积。

2.1.2 空洞卷积

基于图像特点，网络需要有感受野较小的卷积核来提取小尺寸特征，还有感受野较大的卷积核，来提取低像素特征。空洞卷积的感受野可调，能在不增加参数量的同时，保留网络图像的细节信息，有利于提取特征图中不同尺寸特征。解决等倍扩张率序列的空洞卷积采样时丢失大量局部信息问题，使用空洞卷积时，采用混合空洞卷积[14](hybrid dilated convolution)，它是根据扩张率计算公式设计的空洞卷积序列$[r_1,...,r_i,...,r_n]$，实现感受野内信息全覆盖，扩张率小的空洞卷积提取基础信息，扩张率大的空洞卷积提取长距离信息，获取更大感受野范围，又能保持运算量大小不变。公式中 r_i 是第 i 层的膨胀率，M_i 是第 i 层最大膨胀率

$$M_i=\max\left[M_{i+1}-2r_i,M_{i+1}-2(M_{i+1}-r_i),r_i\right]。\tag{7}$$

利用上式计算，混合空洞卷积为连续 3 层卷积核大小均为 3×3，扩张率分别为 1、2、3。ResNet50 网络的 conv4_x 部分由 6 个残差模块堆叠而成，第一个残差模块的输入特征图尺寸为 28×28，其余 5 个残差模块的输入特征图为 14×14。conv5_x 部分由 3 个残差模块堆叠而成，第一个残差模块的输入特征图尺寸为 14×14，其余 2 个残差模块的输入特征图为 7×7，conv5_x 部分的特征图尺寸太小。因此在 conv4_x 部分引入混合空洞

卷积,使用混合空洞卷积序列[1,2],conv4_x部分3×3标准卷积的扩张率序列为[1,2,1,2,1,2]。

基于ResNet50进行改进,一方面在conv2_x和conv3_x部分引入空间注意力机制,使用选择核卷积替换bottleneck残差模块中3×3标准卷积;另一方面在conv4_x部分应用锯齿状混合空洞卷积[1,2,1,2,1,2],即使用卷积核尺寸为3×3、扩张率为2的空洞卷积替换conv4_x部分的第二、第四和第六个bottleneck残差模块的3×3标准卷积,图8展示了改进的ResNet50网络结构。

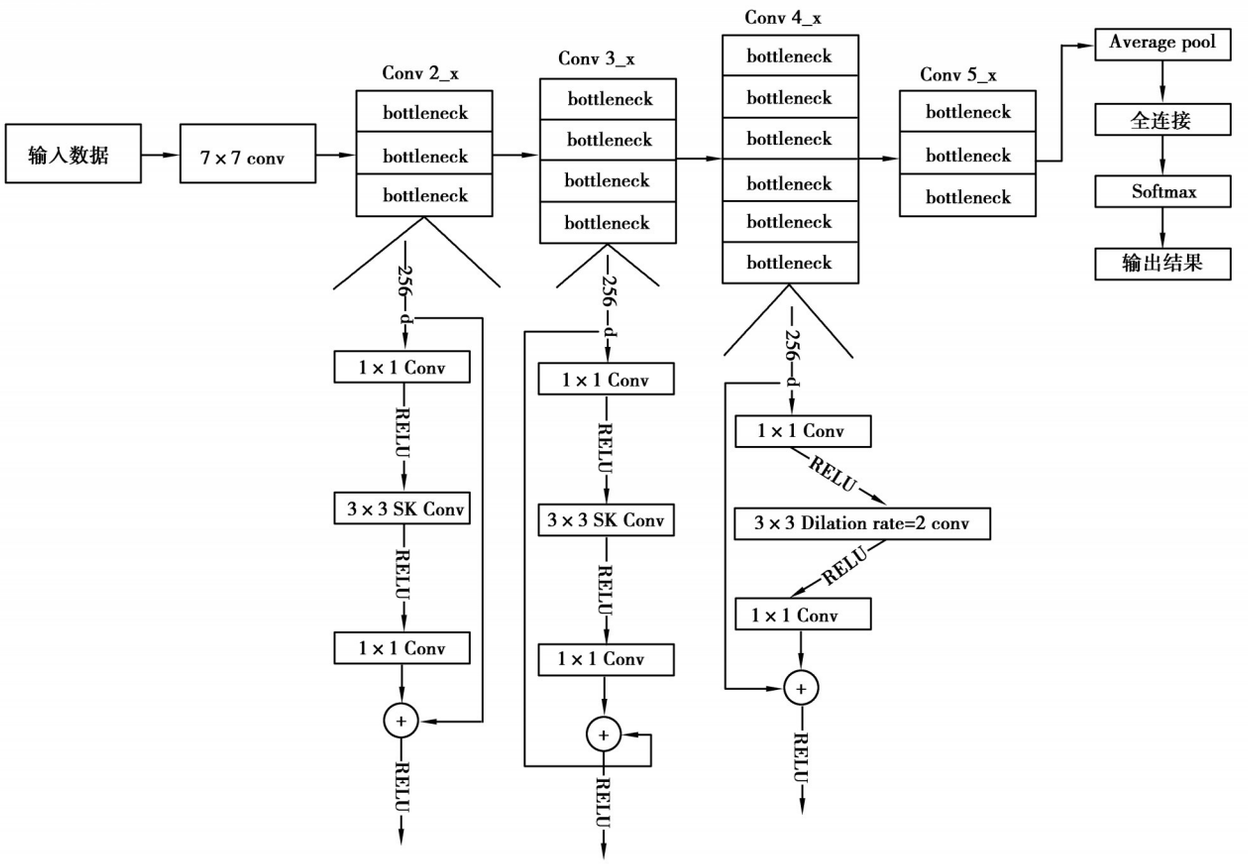


图8　改进的ResNet50的网络结构(AR,Attention ResNet)

Fig. 8　Network structure of the improved ResNet50 (AR, Attention ResNet)

2.2　基于IoU优化的YOLOv3的道路目标边框识别

YOLOv3是由Joseph Redmon和Ali Farhadi提出的,网络的主体框架为Darknet-53结构,共有53个卷积层,代替了YOLOv2中的Darknet-19,与其相比,Darknet-53属于全卷积网络,因为没有最大池化层,下采样操作也是卷积层实现,与其并肩的网络ResNet相比,Darknet-53的卷积核个数、运算量、速度都更强。卷积层、批量归一化层以及LeakyReLU激活函数共同组成Darknet-53中的基本卷积单元DBL[15]。Darknet-53结构图及DBL如图9-10所示(以输入图像尺寸为416×416为例)。Darknet-53的特征提取部分借助了残差网络思想,残差结构如图11所示。YOLOv3网络共使用了5个残差块,对其中的第3、4、5个残差块所提取出的8倍、16倍和32倍下采样特征图进行目标识别。YOLOv3的结构如图12所示(以输入图像尺寸为41。YOLOv3中的定位损失使用差值平方的计算方法,也就是L2损失。但在实际情况中,即使2个目标边界框的重合程度不同,求得的L2损失可能相同,只有2个目标边界框重合程度越高,损失越小,L2损失的弊端因此显现。IoU被广泛使用是因为相比于L2损失,IoU损失能更好反映预测边界框与真实边界框的重合程度,且具有尺度不变性[16-17],即在整个空间中,2个目标边界框在不同尺度大小下可以保持不变,后来也被用到YOLOv3的目标检测方法中,但其也有一些缺点。

	Type	Filters	Size	OutPut
	Convolutional	32	3 × 3	416 × 416
	Convolutional	64	3 × 3/2	208 × 208
1 ×	Convolutional	32	1 × 1	
	Convolutional	64	3 × 3	
	Convolutional			208 × 208
	Convolutional	128	3 × 3/2	104 × 104
2 ×	Convolutional	64	1 × 1	
	Convolutional	128	3 × 3	
	Residual			104 × 104
	Convolutional	256	3 × 3/2	52 × 52
8 ×	Convolutional	128	1 × 1	
	Convolutional	256	3 × 3	
	Residual			52 × 52
	Convolutional	512	3 × 3/2	26 × 26
8 ×	Convolutional	256	1 × 1	
	Convolutional	512	3 × 3	
	Residual			26 × 26
	Convolutional	1 024	3 × 3/2	13 × 13
4 ×	Convolutional	256	1 × 1	
	Convolutional	1 024	3 × 3	
	Residual			13 × 13
	Avg pool		Global	
	Connected		1 000	
	Softmax			

图 9 Darknet-53 结构

Fig. 9 Darknet-53 structure

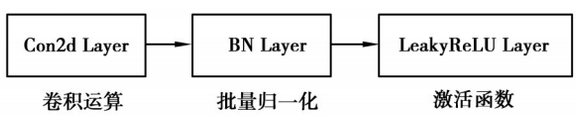


图 10 基本卷积单元 DBL

Fig. 10 Basic convolution unit DBL

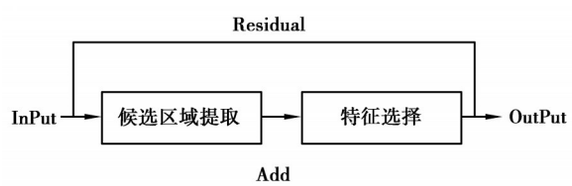


图 11 残差结构图

Fig. 11 Residual structure diagram

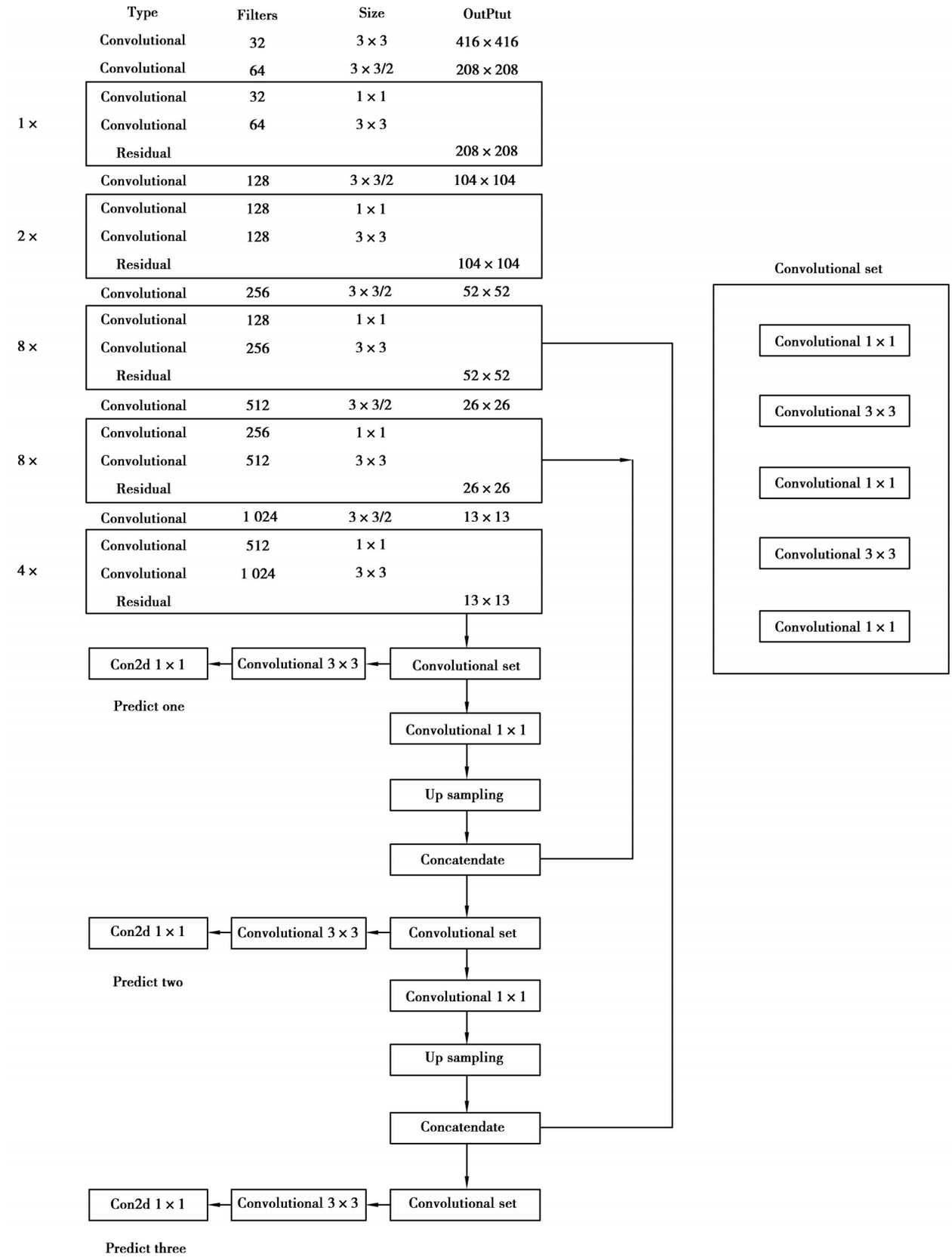


图 12　YOLOv3 网络结构

Fig. 12　YOLOv3 network structure

1)IoU对于预测边界框和真实框的位置要求较高,只有当2个框有交集时,其计算公式才奏效,对于完全没有相交的2个框来说,IoU损失计算为0,无法将损失反馈到神经网络中,没有梯度回传,就无法进行学习训练,影响更新网络权重,使网络一直处在局部最优值附近,始终无法收敛到全局最优。

2)在IoU损失计算过程中,无法判定预测边界框和真实边界框的关系,如方向关系,即当目标物和检测框呈现不同水平方向,夹角无法进行检测。

针对IoU出现的问题,文中引入GIoU损失函数,假定针对2个矩形A和B,能够找到2个矩形的最小外接矩形C。GIoU计算方法如下式

$$\mathrm{GIoU}=\mathrm{IoU}-\frac{A^c-U}{A^c},L_{\mathrm{GIoU}}=1-\mathrm{GIoU}, \tag{8}$$

式中:IoU为预测边界框和真实边界框的交并比,A^c为2框的最小外接矩形C的面积,U为2框并集的面积,模型为ARIY3(Attention-ResNet50-IoU- YOLOv3)。

2.3 点云数据与RGB数据信息融合模型

由于16线激光雷达点云数目特别稀少,导致反射率不太稳定,因对点云数目过少、或未识别出的模糊数据,在摄像头的像素点与激光雷达的点云标定之后,与16线激光雷达和相机传输回来的信息相互融合,获取目标物体的信息,实现目标跟踪。

为了将ResNet50输出的特征融合到原有点云特征提高点云稀疏目标的检测精度,分别使用2个卷积核大小为1×1的卷积层,将图像特征分别压缩到1×1×p和1×1×q尺寸。YOLOv3与激光雷达网络与3D边界框估计网络组成一个整体,进行端到端训练,为后二者的任务筛选出最具价值信息,本文模型如图13所示。

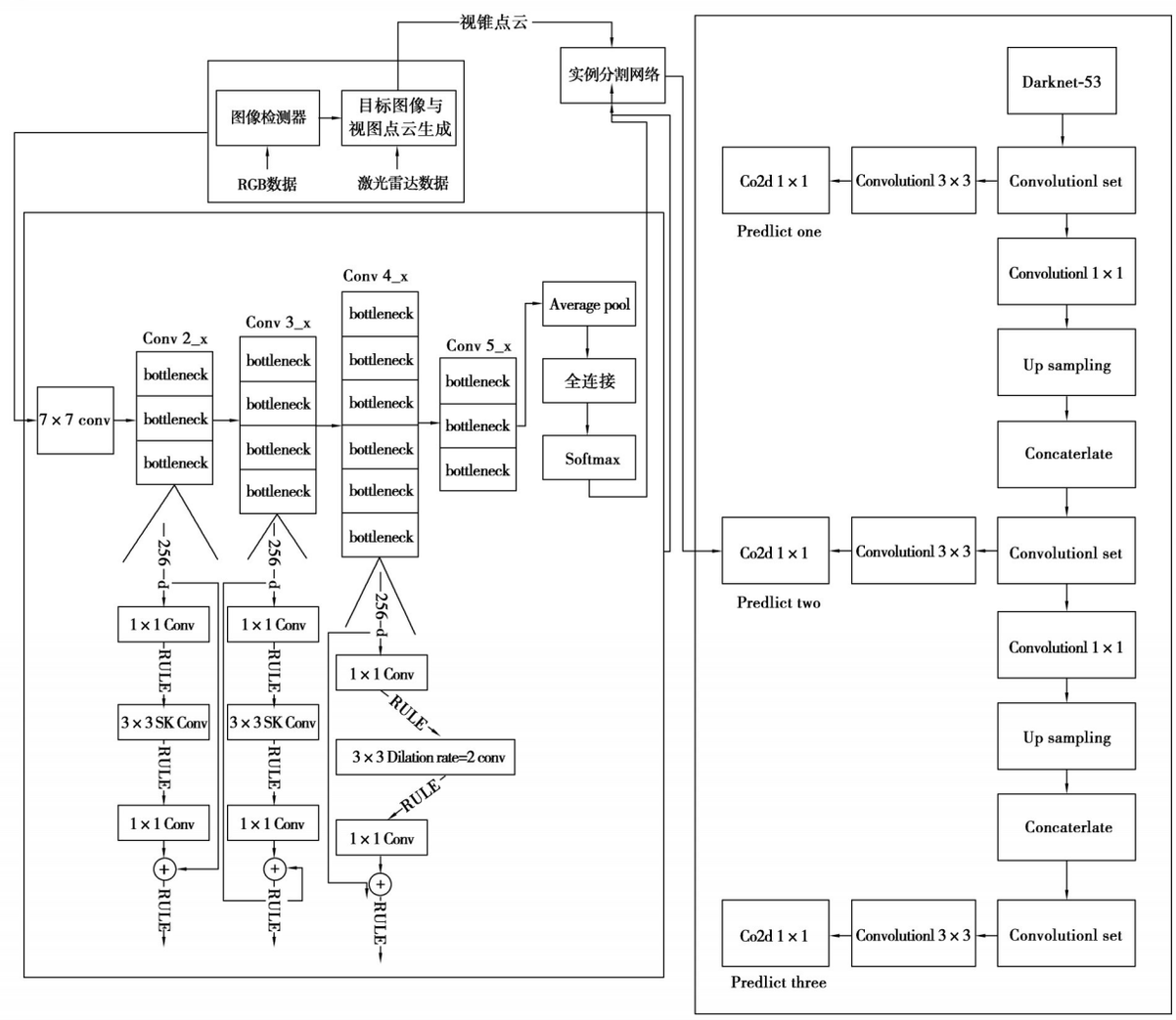


图13 LiDAR-RGB-ARIY3

Fig. 13 LiDAR-RGB-ARIY3

3　实　验

3.1　实验设置

文中使用的训练和测试数据基于KITTI[17]目标检测数据集中的激光点云和左彩色相机数据,其中激光点云处理后全部进行图像化编码,构建为图像化点云数据集。笔者将该数据集的7 481张训练图像作为实验数据,并根据需求预处理数据集原有的标签信息,处理后的整个数据集按照训练集:验证集:测试集=8:1:1的比例进行随机划分,划分后的训练、验证和测试数据集大小分别为5 984、748和749,数据集样本如图14所示。

图14　数据集样本数据示意

Fig. 14　Schematic representation of sample data in the data set

实验使用的操作系统为Ubuntu16.04,GPU为NvidiaRTX2080Ti,显存为11G。实验采用Pytorch1 5.0框架对模型进行搭建、训练和测试,Python版本为3.7,CUDA版本为10.1。在训练阶段,根据显存大小将batchsize设置为8,每个批次中的输入图像尺寸都被固定至512×512大小。动量配置为0.937,权重衰减配置为0.000 5,初始学习率为10^{-3}。实验发现,当程序运行到60 000代之后,损失值出现震荡不再下降,因此在第60 000代将学习率设置为原来的0.1实现损失值继续小范围下降,达到更好拟合效果。下图为训练过程中的损失函数收敛曲线,从图15中看出,训练次数达到100 000次时损失函数收敛曲线趋于平缓。

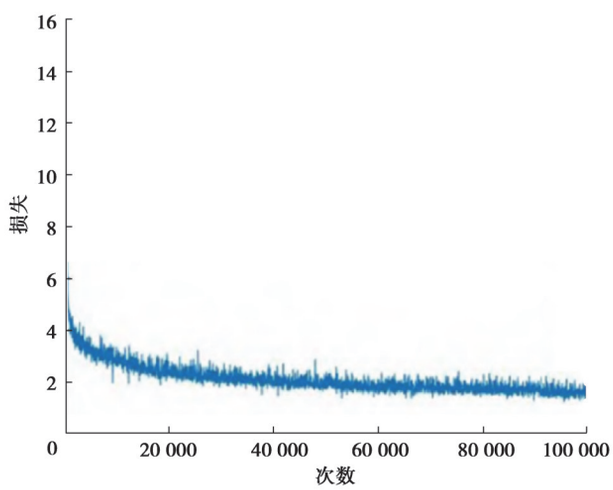


图15　训练迭代次数与损失值的关系

Fig. 15　Relationship between the number of training iterations and the loss value

研究使用目标检测任务中常用的指标P-R曲线和mAP(mean average precision)平均精度2项指标对所提出的模型进行评价。在绘制PR曲线时,首先通过真正例(true positive,TP),真反例(true negative,TN),假正例(false positive,FP),假反例(false negative,FN)计算准确率Precision和召回率Recall,公式如下

$$\text{precision}=\frac{TP}{TP+FP},\text{Recall}=\frac{TP}{TP+FN}。\tag{9}$$

针对某一类别,以召回率为横轴,以准确率为纵轴可以绘制P-R曲线,曲线所包含的面积即为该类别的AP。mAP则是对这多种类别的AP值求平均所得。AP值代表模型对某一类目标的检测效果,mAP则代表了

对所有类别的检测效果，值越大，检测效果越好。实验设置初始IoU阈值为0.5，使用GIoU检测预测框与真实框的交并比划分样本。

3.2 实验分析

3.2.1 消融实验

为了验证利用LiDAR-RGB-ARIY3进行特征级融合的效果，研究采用图像化点云数据和RGB图像数据，在ARIY3架构下分别训练了3种模型，即ARIY3（RGB）、ARIY3（LiDAR）和LiDAR-RGB-ARIY3。通过对比单数据模型、融合数据模型以及不同融合方式，评估各模型性能。其中，ARIY3（RGB、LiDAR）是通过特征级融合训练得到的模型，它将2种数据直接进行通道级联，将联合的特征输入到ARIY3进行训练。这一方法旨在充分发挥LiDAR和RGB数据在特征级上的互补性，提高模型的性能和泛化能力。通过这一对比实验，可以深入了解不同数据和融合方式对最终模型性能的影响，为LiDAR与RGB数据融合的有效性提供实证支持。

网络设定推理的目标得分阈值为0.24，NMS阈值为0.5，计算AP和mAP时的IOU设定为50%，对训练好的模型在测试集上进行对比实验，结果如表1所示。

表1 不同融合方法性能对比

Table 1 Performance comparison of different fusion methods

Model	白天mAp	夜晚mAp
ARIY3（RGB）	0.77	0.51
ARIY3（LiDAR）	0.75	0.71
ResNet-YOLOv3(RGB、LiDAR)	0.81	0.79
ARIY3（RGB、LiDAR）	0.92	0.91

对比LiDAR-RGB-ARIY3、ResNet-YOLOv3(RGB、LiDAR)、ARIY3（RGB）和ARIY3（LiDAR）可以看出，相对于单数据模型，基于激光点云和RGB图像的融合模型具有更好检测效果。在白天视线较好条件下，ResNet-YOLOv3(RGB、LiDAR)比ARIY3（RGB）和ARIY3（LiDAR）分别提升0.04和0.06。在黑夜视线较差条件下，ResNet-YOLOv3(RGB、LiDAR)比ARIY3（RGB）和ARIY3（LiDAR）分别提升0.28和0.08。而在白天视线较好条件下，ARIY3（RGB、LiDAR）比ResNet-YOLOv3(RGB、LiDAR)的mAP提升0.11，在黑夜视线较差条件下，ARIY3（RGB、LiDAR）比ResNet-YOLOv3(RGB、LiDAR)的mAP提升0.12。实验结果表明，融合特征对目标具有更强表征性，多模态融合无论白天还是夜晚，均有利于提高检测网络性能。其中多模态特征融合对于网络提升效果较为明显，特别是在低照度场景下。同时，ARIY3（RGB、LiDAR）比ResNet-YOLOv3(RGB、LiDAR)的mAP有所提升，实验结果表明所提出的目标识别方法在光照变化的场景依然表现出较好鲁棒性。

在当前的配置环境下，完成整个KITTI训练集上双模态深度学习网络的100 000次迭代大约需要15 h。损失函数（loss）在网络模型训练过程中的演变如图16所示。图中绿色和红色虚线分别代表训练单模态的雷达激光图像目标识别网络和可见光图像目标识别网络的损失，蓝色实线表示双模态目标识别网络在原ResNet-YOLOv3后进行融合的模型训练损失，而黑色实线则表示双模态目标识别网络在LiDAR-RGB-ARIY3进行融合的模型训练损失。通过观察图16，可以得知在经过100 000次迭代后，所有模型表现出良好的收敛效果。

通过局部放大图中的细节，相较于单模态网络训练，多模态目标识别网络训练损失变化更加平缓，模型更快收敛。在给定的训练迭代次数内，多模态网络在学习目标识别任务上表现出更高的效率和稳定性。这些结果进一步验证了双模态深度学习网络在LiDAR和RGB数据融合方面的优越性。

图17所示是LiDAR-RGB-ARIY3在验证集数据上的检测可视化结果，图中红色框为真值框，蓝色框为网络的预测输出，框中线条代表检测框中心延伸出的方向向量。从图中标记的目标看出：虽然目标在图像视角中像素面积小，以至于真值都未对其进行标注，但网络通过融合点云和图像特征将其检测出来，表明使用

多模态传感器融合对遮挡、距离较远目标识别具有一定优势。

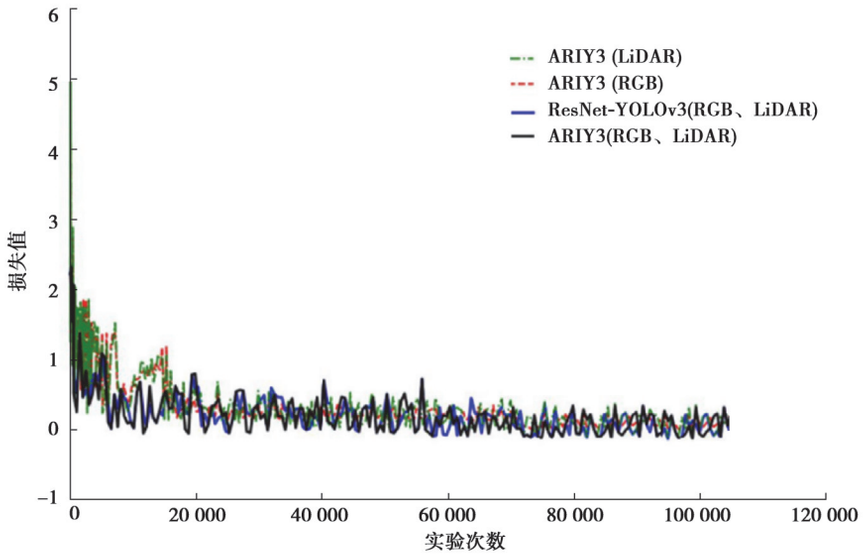


图 16　消融实验双模态网络损失曲线

Fig. 16　Bimodal network loss curve of ablation experiment

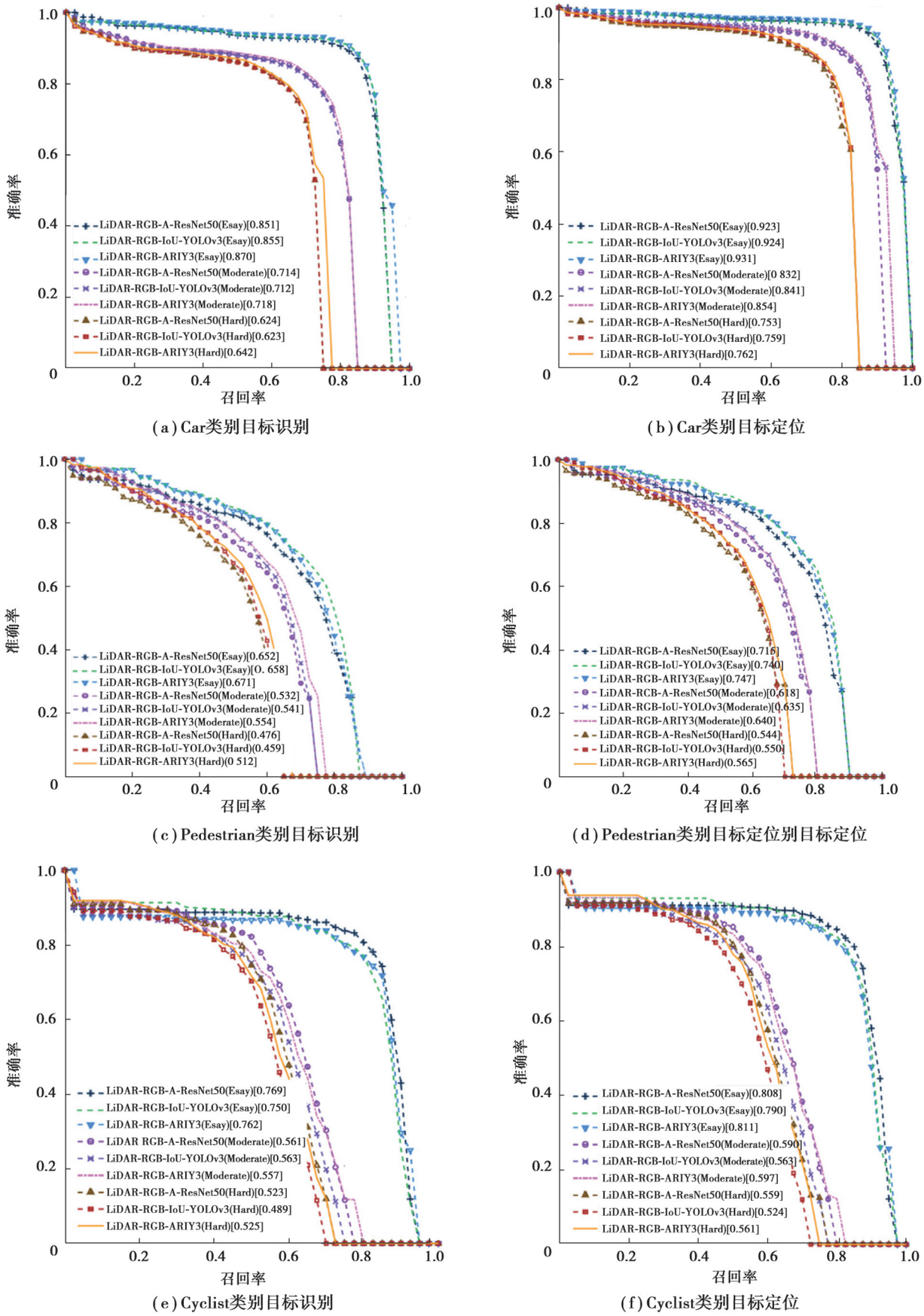
图 17　验证集测试可视化结果

Fig. 17　Verification set test visualization results

KITTI的数据集根据目标的检测框大小、受遮挡情况和在视野中被截断面积,对目标识别的难易程度进行划分,划分为简单(Easy)、适中(Moderate)和困难(Hard)。实验按照目标识别的难易程度,对检测性能进一步评估。

将 LiDAR-RGB-ARIY3 以及提出的 LiDAR-RGB-A-ResNet50(与 LiDAR-RGB-ARIY3 相比,仅优化 ResNet,不优化 YOLOv3 的多模态信息融合模型)与 LiDAR-RGB-IoU-YOLOv3(与 LiDAR-RGB-ARIY3 相比,仅优化 YOLOv3,不优化 ResNet 的多模态信息融合模型)在 KITTI 数据集上分别进行 3 种目标类别的 2 种挑战后,实验结果如图 18 所示的 P-R 曲线图。

从图中可以看到,LiDAR-RGB-ARIY3 在 Car 类别与 Pedestrian 类别都获得了显著提升,同时 Cyclist 类别总体来说相差细微。从图(c)和(d)来看,LiDAR-RGB-ARIY3 在 Pedestrian 类别的目标识别上远超过 LiDAR-RGB-A-ResNet50 以及 LiDAR-RGB-IoU-YOLOv3,仅在召回率较低时保持与原始方法的较大优势不同,LiDAR-RGB-ARIY3 在所有召回率位置上取得显著优势。对于 Pedestrian 类别的目标定位,研究提出的 2 种方法对 LiDAR-RGB-A-ResNet50 及 LiDAR-RGB-IoU-YOLOv3 都取得了显著优势,其中引入通道注意力机制使 LiDAR-RGB-A-ResNet50 在前一章方法的效果上继续扩大优势。从图 18(a)和(b)来看,



(a) Car类别目标识别　(b) Car类别目标定位

(c) Pedestrian类别目标识别　(d) Pedestrian类别目标定位别目标定位

(e) Cyclist类别目标识别　(f) Cyclist类别目标定位

图 18　消融实验对比结果

Fig. 18　Comparison results of ablation experiments

LiDAR-RGB-ARIY3 相对 LiDAR-RGB-A-ResNet50 以及 LiDAR-RGB-IoU-YOLOv3 方法同样获得提升。除了在简单难度以外,LiDAR-RGB-ARIY3 在所有其它项中均取得领先。对于通道注意力的引入对于尺寸较小以及存在遮挡的目标检测具有明显提升效果。

从图(e)和(f)来看,LiDAR-RGB-A-ResNet50 以及 LiDAR-RGB-IoU-YOLOv3 在 Cyclist 类别的检测与定位方面差距细微。同时,发现原始方法的优势出现在召回率较高时,在召回率较低时,LiDAR-RGB-ARIY3 则有明显优势,这意味着 LiDAR-RGB-ARIY3 对于其检测的高置信度目标有更高的准确率。

3.2.2　对比实验

车辆的点云与图像区域如图 19 所示。

图 19　车辆对应的点云与图像区域

Fig. 19　Point cloud and image area corresponding to the vehicle

为了评估所提出的多模态特征融合目标识别网络性能,笔者设计对比实验。实验中将该方法与 FasterRCNN、OFTNet 和 VoxelNet 在 2 种光照环境下的性能展开对比。表 2 展现不同方法在 KITTI 数据集上目标识别的对比结果。

表 2　不同方法目标识别对比

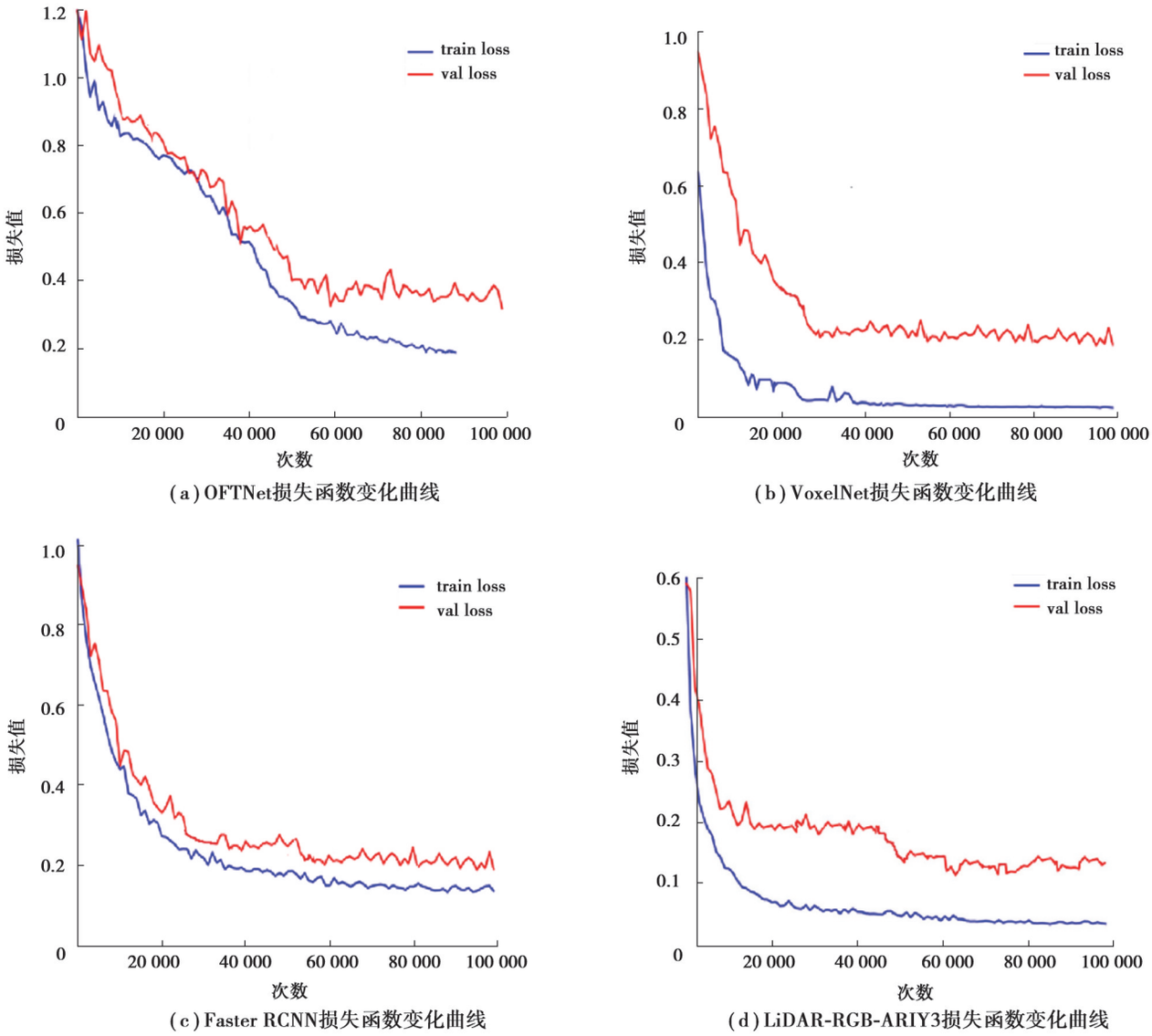
Table 4-2　Comparison of different methods of target identification

Model	白天 mAp	夜晚 mAp
OFTNet[18]	0.87	0.82
VoxelNet[19]	0.90	0.86
Faster RCNN	0.88	0.85
LiDAR-RGB-ARIY3	0.92	0.91

从白天对比实验结果看出,相较与 OFTNet 、VoxelNet 和 FasterRCNN 网络,提出的多模态特征融合检测方法在 *AP* 指标上均有提升,尤其是在 Faster RCNN 模式上,mAP 指标提升 0.04,较为明显。该对比实验证明方法在光照良好场景具有较好的检测性能。从夜间对比实验结果可以看出,相较与 OFTNet 、VoxelNet 和 FasterRCNN 网络,提出的多模态特征融合检测方法在 *AP* 指标上均有提升,提升幅度最大可达到 0.09,较为明显。该对比实验证明了该方法在低照度场景具有较好的检测性能。

各模型训练和验证过程中的损失函数变化曲线如图 20 所示,每个 Epoch 进行。由图可知,提出的 LiDAR-RGB-ARIY3 模型训练集损失函数和验证集损失函数耗能最低,表明模型中每个样本预测值和真实值的差最小,所建立的模型提供的结果最好[20]。

综上所述,笔者提出的自适应融合网络 LiDAR-RGB-ARIY3 与常见的基于点云、基于多模态融合的网络相比,检测精度与速度有一定优势,实现精度与速度的平衡,图 21 为可视化结果。



(a) OFTNet损失函数变化曲线

(b) VoxelNet损失函数变化曲线

(c) Faster RCNN损失函数变化曲线

(d) LiDAR-RGB-ARIY3损失函数变化曲线

图 20 损失函数变化曲线

Fig. 20 Loss function change curve

图 21 可视化结果展示

Fig. 21 Visualization results display

4 结　论

研究提出一种基于激光雷达和视觉传感器信息融合的无人驾驶中目标识别算法。该算法主要包括以下几个改进方面:

1)利用摄像头的视觉方案识别目标物体图片,图片经过预处理,传入卷积神经网络ResNet50进行特征提取,使用yolov3改进算法得到物体的类别与物体框位置信息。

2)使用注意力机制对ResNet50进行改进,集中在网络的特征提取部分。使用优化的IoU对YOLOv3模型的目标边框提取进行完善。

3)利用激光雷达进行地面点距离标定,将像素点与激光雷达的标定点进行对应,对点云数据和图像数据进行时间、空间同步,得到激光雷达和相机数据之间的转换关系,找到同一时刻激光点云数据和图像中对应的像素点,确保激光雷达识别出的物体与相机识别的物体是同一时刻同一物体。

该算法目的是解决无人驾驶环境下运动目标检测问题,通过多源数据融合的方式提高目标检测的准确率。该算法在进行数据融合时,没有进行时间、空间同步,这个过程可能环境因素会影响数据的准确性,如天气、光照等。未来考虑加入时间、空间同步方法,以提高数据融合的准确性。

参考文献

[1] 熊璐,吴建峰,邢星宇,等.自动驾驶汽车行驶风险评估方法综述[J/OL].汽车工程学报:1-15 [2023-04-28]. 网址:http://kns.cnki.net/kcms/detail/50.1206.U.20230425.0916.002.html
Xiong L, Wu J F, Xing X Y, et al. Review of automatic driving vehicle driving risk assessment methods[J/OL]. Automotive Engineering Journal: 1-15[2023-04-28].http://kns.cnki.net/kcms/detail/50.1206.U.20230425.0916.002. html(in Chinese)

[2] Nan Y L,Zhang H C, Zeng Y. Intelligent detection of Multi-Class pitaya fruits in target picking row based on WGB-YOLO network[J]. Computers and Electronics in Agriculture,2023,208: 107780.

[3] Li J R, Cai R Y, Tan Y, et al. Automatic detection of actual water depth of urban floods from social media images[J]. Measurement,2023,216: 1-19.

[4] Vora S, Lang A H, Helou B, et al. Pointpainting: sequential fusion for 3d object detection[C]//Proc of Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020: 4604-4612.

[5] Ren S, He K, Girshick R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[J]. IEEE transactions on pattern analysis and machine intelligence, 2016, 39(6):1137-1149.

[6] Ku J, Mozifian M, Lee J, et al. Joint 3d proposal generation and object detection from view aggregation【C]//Proc of 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Madrid: IEEE, 2018, 1-8.

[7] Botha F. Data fusion of radar and stereo vision for detection and tracking of moving objects[C]//Pattern Recognition Association of South Africa & Robotics & Mechatronics International Conference. Bloemfontein: IEEE, 2017.

[8] Li Y, Ma L, Zhong Z, et al. Deep learning for lidar point clouds in autonomous driving: a review [J]. IEEE Transactions on Neural Networks and Learning Systems, 2020(99):1-21.

[9] Wang Y X, Xu S S, Li W B, et al. Identification and location of grapevine sucker based on information fusion of 2D laser scanner and machine vision)[J]. International Journal of Agricultural and Biological Engineering, 2017,10(2), 84-93.

[10] Barrientos A, Garzón M, Fotiadis P E .Human detection from a mobile robot using fusion of laser and vision information[J]. Sensors,2013,13(9):11603-11635.

[11] Huang Y, Xiao Y, Wang P, et al.A seam-tracking laser welding platform with 3D and 2D visual information fusion vision sensor system[J].The International Journal of Advanced Manufacturing Technology,2013,67(1-4):415-426.

[12] Ajayi O G, Ashi J, Guda B. Performance evaluation of YOLO v5 model for automatic crop and weed classification on UAV images[J]. Smart Agricultural Technology,2023,5: 1-10.

[13] He K M, Zhang X, Ren S, et al. Deep Residual Learning for Image Recognition[C]//2016 IEEE Conference on Computer

Vision and Pattern Recognition, 2016: 770-778.

[14] Li Z,Xu B L,Wu D, et al. A YOLO-GGCNN based grasping framework for mobile robots in unknown environments[J]. Expert Systems With Applications,2023, 225: 1-14.

[15] Zhao C, Shu X, Yan X, et al. RDD-YOLO: a modified YOLO for detection of steel surface defects[J]. Measurement,2023,214: 1-12

[16] 邹承明,薛榕刚.GIoU和Focal loss融合的YOLOv3目标检测算法[J].计算机工程与应用,2020,56(24):214-222.
Zou C M, Xue R G. Improved YOLOv3 object detection algorithm:combining GIoU and Focal loss[J]. Computer Engineering and Applications, 2020, 56(24):214-222.(in Chinese) .

[17] Geiger A, Lenz P, Urtasun R. Are we ready for autonomous driving? the kitti vision benchmark suite[C]//2012 IEEE conference on computer vision and pattern recognition. IEEE, 2012: 3354-3361.

[18] Roddick T, Kendall A, Cipolla R. Orthographic feature transform for monocular 3d object detection[J]. arXiv preprint arXiv: 1811.08188, 2018.

[19] Zhou Y, Tuzel O. Voxelnet: end-to-end learning for point cloud based 3d object detection[C]//Proc of Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018, 4490-4499.

[20] 吴喆.基于深度学习的动态背景下船舶检测和跟踪的研究[D].宜昌：中国三峡大学,2019.
Wu Z. Research on ship detection and tracking in dynamic background based on deep learning[D]. Yichang: China Three Gorges University, 2019.(in Chinese)

（编辑 侯 湘）

（上接第21页）

[23] Zhang S W, Han B, Sun Y H, et al. Microplastics influence the adsorption and desorption characteristics of Cd in an agricultural soil[J]. Journal of Hazardous Materials, 2020, 388: 121775.

[24] 杜海玲, 张迎霜, 王晖, 等 . 不同微塑料对亚甲基蓝的吸附行为[J]. 环境化学, 2022, 41(9): 2803-2812.
Du H L, Zhang Y S, Wang H, et al. Adsorption behavior of methylene blue by different microplastics[J]. Environmental Chemistry, 2022, 41(9): 2803-2812.(in Chinese)

[25] Queiroz H M, Ferreira T O, Barcellos D, et al. From sinks to sources: the role of Fe oxyhydroxide transformations on phosphorus dynamics in estuarine soils[J]. Journal of Environmental Management, 2021, 278: 111575.

[26] Gérard F. Clay minerals, iron/aluminum oxides, and their contribution to phosphate sorption in soils: a myth revisited[J]. Geoderma, 2016, 262: 213-226.

[27] Andersson K O, Tighe M K, Guppy C N, et al. The release of phosphorus in alkaline vertic soils as influenced by pH and by anion and cation sinks[J]. Geoderma, 2016, 264: 17-27.

[28] Wang Z H, Guo H Y, Shen F, et al. Biochar produced from oak sawdust by Lanthanum (La)-involved pyrolysis for adsorption of ammonium (NH_4^+), nitrate (NO_3^-), and phosphate (PO_4^{3-})[J]. Chemosphere, 2015, 119: 646-653.

[29] Du C, Ren X Y, Zhang L A, et al. Adsorption characteristics of phosphorus onto soils from water level fluctuation zones of the Danjiangkou Reservoir[J]. CLEAN - Soil, Air, Water, 2016, 44(8): 975-983.

[30] Liu L, Song J, Zhang M, et al. Aggregation and deposition kinetics of polystyrene microplastics and nanoplastics in aquatic environment[J]. Bulletin of Environmental Contamination and Toxicology, 2021, 107(4): 741-747.

[31] Johansen M P, Cresswell T, Davis J, et al. Biofilm-enhanced adsorption of strong and weak cations onto different microplastic sample types: use of spectroscopy, microscopy and radiotracer methods[J]. Water Research, 2019, 158: 392-400.

[32] Li S, Yang M X, Wang H, et al. Adsorption of microplastics on aquifer media: effects of the action time, initial concentration, ionic strength, ionic types and dissolved organic matter[J]. Environmental Pollution, 2022, 308: 119482.

[33] Wang F Y, Yang W W, Cheng P, et al. Adsorption characteristics of cadmium onto microplastics from aqueous solutions[J]. Chemosphere, 2019, 235: 1073-1080.

（编辑 郑 洁）