一种基于 YOLOv5 的轻量型行人检测方法

A Lightweight Pedestrian Detection Method Based on YOLOv5

王 亮 张 超 (西安工程大学电子信息学院,陕西 西安 710048)

摘要:基于卷积神经网络的行人检测方法对行人检测性能有了很大的提升,但当前行人检测方法计算量大,占用了较多的计算机资源。为了解决这种问题,提出了一种改进的 YOLOv5 行人检测算法。该算法在保证检测精度不变、减小权重大小、提升检测速度的情况下,在原有的 YOLOv5 网络基础上引入轻量级卷积模块 Ghost 卷积,并且为了提高检测精度,在主干网络中加入轻量注意力机制 ECA。为了进一步提高检测精度,将原有的特征融合网络 PAN+FPN 结构替换为加权双向金字塔结构 BiFPN。通过实验结果表明,经过引入和替换模块后,模型网络精度保持不变,模型大小减小了约 2.13 倍,浮点型计算量减少了约 2.51 倍,检测速度(FPS)提高了约 1.67 倍。

关键词:深度学习;YOLO;行人检测;轻量级网络;注意力机制

Abstract: Pedestrian detection method based on convolutional neural network has greatly improved the performance of pedestrian detection, but the current pedestrian detection method has a large amount of computation and occupies more computer resources. In order to solve this problem, an improved YOLOv5 pedestrian detection algorithm is proposedin this paper. In this algorithm, the lightweight convolution module Ghost convolution is introduced on the basis of the original YOLOv5 network under the condition of keeping the detection accuracy unchanged, reducing the weight and improving the detection speed. In order to improve the detection accuracy, the lightweight attention mechanism ECA is added to the backbone network. The original feature fusion network PAN+FPN is replaced by weighted bidirectional pyramid structure BiFPN.

Keywords: deep learning, YOLO, pedestrian detection, lightweight network, mechanism of attention

随着人工智能技术的快速发展，基于深度学习的目标检测算法已经应用到多个领域之中,例如:遥感技术、医学成像、无人驾驶等。其中有很多场景都需要实时反馈,因此检测精度及速度都有高要求。由于目标检测算法在精度与速度方面都有了很大的进步，但其网络结构越发复杂，当模型被部署到移动端设备时,由于移动端设备计算能力弱,所以模型的执行速度会大幅度下降。因此对模型进行轻量化的研究有着重要的意义。

目前基于深度学习的目标检测算法主要分为双阶段和单阶段两种检测算法,其中双阶段检测算法以基于候选框[1]的方法为主,将检测算法分为检测和识别两个阶段。此类检测算法的检测精度较高,但是检测速度较慢。如:RCNN[2]、Fast RCNN[3]、Faster RCNN[4]和 Mask RCNN[5]等;单阶段检测算法以基于回归[6]的方式为主,去掉了区域选择算法。此类算法检测精度较低,但检测速度很快。如 SSD[7]、CenterNet[8]、YOLO[9-12]系列等。在目标检测领域中行人检测只是其中一部分问题，通过卷积神经网络可以从海量样本中选择出特征,使得检测结果更好,这也是基于深度学习的行人检测模型的优点之一。本文提出了一种基于 YOLOv5 网络改进的轻量化行人检测算法。

1 YOLOv5 基本原理

目前 YOLOv5 系列算法已经有 6 个版本，本次实验采用 6.0 版本。YOLOv5 按照网络的深度和宽度将网络模型由小到大分为 YOLOv5s、YOLOv5m、YOLOv5l、YOLOv5x，检测精度和速度呈现递增。本次研究考虑到模型大小问题,采用 YOLOv5s 作为研究对象。

图 1 展示的是 YOLOv5s 的网络结构，包含输入端、Backbone(骨干网络)、Neck(多尺度融合)和 Prediction(检测头)四个部分。

1.1 输入端

输入端为三通道的 RGB 图像，以 640×640×3 为特征大小，采用了 Mosaic 数据增强、自适应图片缩放等技术。首先，Mosaic 数据增强是将 1 张选定的图片和随机的 3 张图片进行随机裁剪,再拼接到一张图上作为训练数据。这样可以丰富图片的背景,而且 4 张图片拼接在一起变相提高了 batch_size,在进行 batch normalization(归一化)的时候也会计算 4 张图片。这样丰富了数据集样本,减少 GPU 使用的数量,使网络具有更好的鲁棒性。其次,YOLOv5s 在输入端还采用自适应锚框计算和自适应图片缩放技术。在针对不同的数据集,都有初始设定长宽的锚框,在初始锚框的基础上将输出预测框和真实框的差值,即以真实边框位置相对于预设框的偏移来计算两者差距，再通过不断迭代,获取最佳的锚框值。在处理不同尺寸图片的时候采用自适应图片缩放,将图片缩放到统一尺寸。

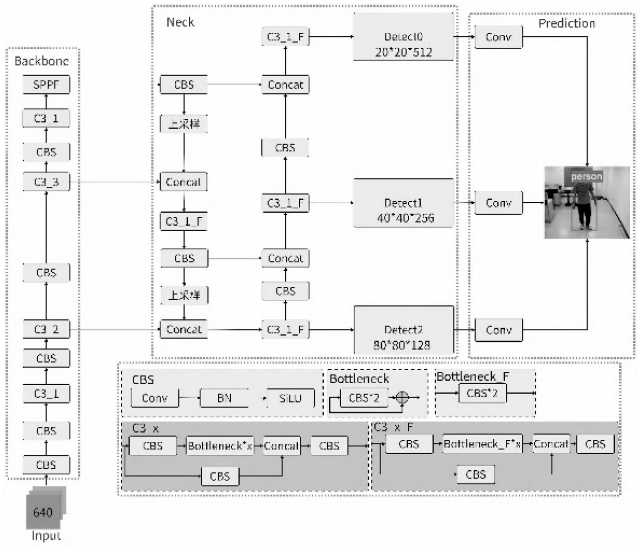

图 1 YOLOv5 网络模型

1.2 Backbone(骨干网络)

Backbone 的主要作用就是进行特征提取，在 YOLOv5s 6.0 版本中骨干网络主要由 CBS 模块、C3 模块和 SPPF 模块组成。其中 CBS 模块包含了卷积层(Conv2d)、归一化层(batch

normalization)和 SiLU 激活函数。C3 模块包含了 3 个标准卷积层以及多个 Bottleneck 层和 Bottlenck_F 层,Bottleneck 层中一路先进行 1×1 卷积将特征图的通道数减小一般,从而减少计算量,再通过 3×3 卷积提取特征,并且将通道数加倍,其输入与输出的通道数不发生改变的。而另外一路通过残差连接,与第一路的输出特征图相加,从而实现特征融合。Bottlenck_F 层与 Bottleneck 层相比少了一路残差连接。SPPF 采用多个小尺寸池化核级联,从而融合不同感受野的特征图、丰富特征图的表达能力的情况下,进一步提高了运行速度。

1.3 Neck(多尺度融合)

Neck 层的主要作用是特征融合,由图 2 所示,它是由 FPN[13]+PAN[14]结构组成,这种搭配是在 YOLOv4 中所使用的。其中 FPN 层网络自顶向下传达语义特征,而 PAN 层则自底向上传达定位特征,两者结合从不同的主干层对不同的监测井进行参数聚合,YOLOv5 相比 YOLOv4,将 YOLOv4 中的普通卷积操作替换为 C3 结构,进一步提高了网络特征融合能力。

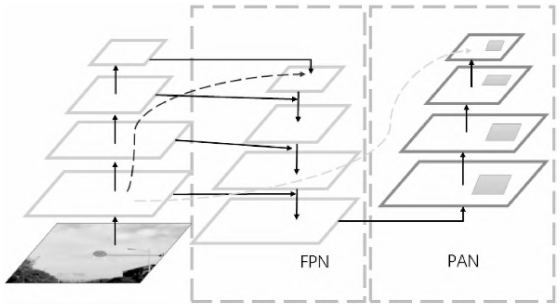

图 2 FPN+PAN 结构

1.4 Prediction(检测头)

Prediction 层包含了 3 个检测层,分别对应 Neck 中得到的 3 种不同尺寸的特征图。YOLOv5 根据特征图的尺寸在这 3 种特征图上划分网格,并且给每种特征图上的网格都预设了 3 个不同宽高比的锚框,用来预测和回归目标。

2 改进的 YOLOv5s 算法

2.1 整体网络架构

本文提出了一种基于 YOLOv5s 的行人检测网络模型,目的是为了减小权重大小且实现精确又高效的行人检测,网络结构如图 3 所示。主要改进点有:在 YOLOv5s 网络中将 CBS 模块替换为 Ghost 卷积模块,将 C3 模块替换为 C3Ghost 模块。

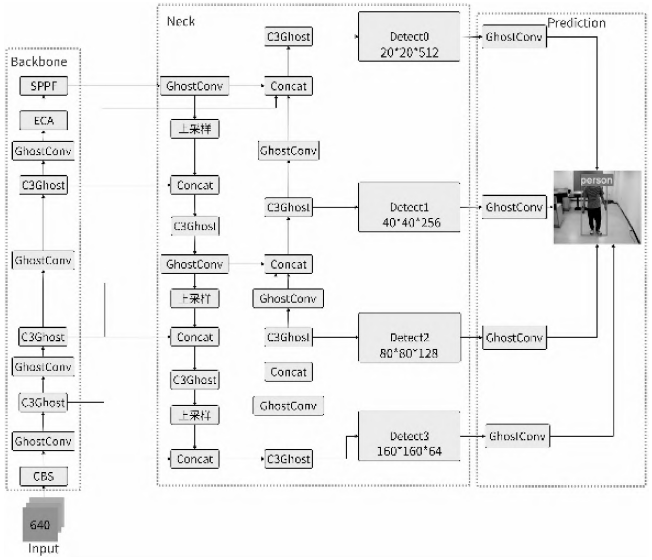

图 3 改进后的 YOLOv5 网络模型

将 ECA 模块引入主干网络中以及在 YOLOv5s 颈部网络层中将 PAN+FPN 模块替换为 BiFPN 模块。其中,Ghost 卷积模块的作用是减小模型的参数以及提高检测速度。ECA 模块的作用是增强特征提取能力,提高行人检测的精度。BiFPN 模块的作用是对于检测不同尺度效率的提升。

2.2 轻量级卷积模块 Ghost 卷积

Ghost 卷积[15]是华为诺亚方舟实验室在 2020 年提出来的一种轻量级卷积模块,其核心思想就是在少量的非线性的卷积得到的特征图基础上,再进行一次线性卷积,从而获取更多的特征图,以此来实现消除冗余特征,获取更加轻量的模型,从而降级计算成本,减少计算资源。图 4 展示了普通卷积和 Ghost 卷积的区别:

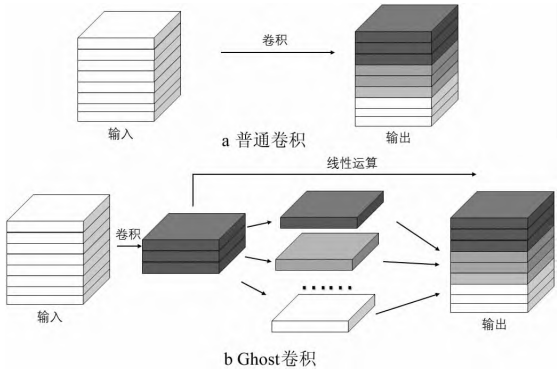

图 4 普通卷积和 Ghost 卷积

图 4a 为普通卷积,图 4b 为 Ghost 卷积。由于普通卷积模块为了高效提取特征,组成多层卷积计算得到特征图,但卷积核与通道数的增加就产生了大量冗余信息,从而导致计算成本增加。与之相比,Ghost 卷积将普通卷积操作分为两部分:第一步,使用少量卷积核进行卷积操作生成一部分特征图;第二步,采用廉价的线性运算生成更多的特征图,然后将不同的特征图连接在一起组成新的输出。

假设卷积核数量为 n,输入特征图大小为 $c\cdot h\cdot w$,分别为输入通道、特征图高和宽,经过一次卷积后为 $n\cdot h'\cdot w'$,卷积核大小为 $k\cdot k$,线性变换卷积核大小为 d,经过 s 次变换。可以推理出普通卷积的计算量为 $n\cdot h'\cdot w'\cdot c\cdot k\cdot k$。Ghost 卷积的计算量为 $\frac{n}{s}\cdot h'\cdot w'\cdot c\cdot k\cdot k+(s-1)\cdot\frac{n}{s}\cdot h'\cdot w'\cdot d\cdot d$。则两者之比为:

$$r_s=\frac{n\cdot h'\cdot w'\cdot c\cdot k\cdot k}{\frac{n}{s}\cdot h'\cdot w'\cdot c\cdot k\cdot k+(s-1)\cdot\frac{n}{s}\cdot h'\cdot w'\cdot d\cdot d}$$
$$=\frac{c\cdot k\cdot k}{\frac{1}{s}\cdot c\cdot k\cdot k+\frac{s-1}{s}\cdot d\cdot d}\approx\frac{s\cdot c}{s+c-1}\approx s \tag{1}$$

式(1)中,$\frac{n}{s}$是第一次变换时的输出通道数,$s-1$ 是因为恒等映射不需要进行计算,但它也算做第二变换中的一部分。从最后结果可以看出,普通卷积的计算量和参数量约为 Ghost 卷积的 s 倍,所以在 YOLOv5 模型中引入 Ghost 卷积可以大幅度降低计算成本和模型的参数量,从而形成新的网络结构——YOLOv5-Ghost。

2.3 ECA 注意力机制

提升网络特征提取性能的方式有很多,注意力机制便是其中之一,因其具有即插即用、有效增强信息的特点被广泛应用于深度学习目标检测领域[16]。SENet(Squeeze-and-Excitation Networks)通道注意力机制是 2018 年由文献[17]提出的,其建立了特征图中的空间相关性,但 SENet 采用的降维操作会对通道注意力的预测产生负面影响,且获取依赖关系效率低且不必

要。基于此,文献[18]在 2020 年提出的一种轻量级高效通道注意力模块 ECA (Efficient Channel Attention Module),避免了降维,有效地实现了跨通道交互,即在 SENet 中使用的全连接层 FC 替换为一维卷积的形式,有效地减少了参数计算量并且对性能有了一定的提升。因此,本文将其应用于 YOLOv5s-Ghost 的主干网络中,对特征提取进行进一步提升。图 5 为 ECA 结构图:

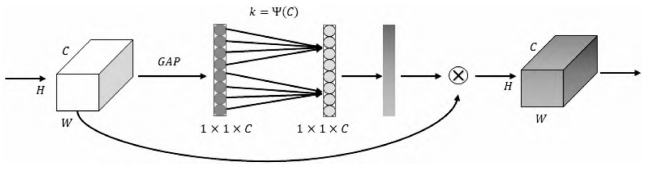

图 5 ECA 结构图

图 5 中,输入特征图的维度为 $C\cdot H\cdot W$,对输入特征图进行空间特征压缩,使用全局平均池化 GAP,得到 1×1×C 的特征图。然后对压缩后的特征图进行通道特征学习,从而实现通过 1×1 卷积,学习不同通道之间的重要性,此时输出的维度为 1×1×C。其中 C 代表通道数,k 代表跨通道交互的邻居数也代表 1×1 卷积的卷积核数。最后通过通道注意力结合,将通道注意力的特征图 1×1×C 及原始输入特征图 $C\cdot H\cdot W$,进行逐通道乘,最终输出具有通道注意力的特征图。其中 k 与 C 之间存在一种映射关系如式(2)所示:

$$k=\Psi(C)=\left|\frac{\log_2(C)}{\gamma}+\frac{b}{\gamma}\right|_{odd} \tag{2}$$

式(2)中 $\left|\frac{\log_2(C)}{\gamma}+\frac{b}{\gamma}\right|_{odd}$ 表示距离最近的奇数 t,γ 和 b 分别设为 2 和 1。

2.4 BiFPN 加权双向金字塔结构

在 YOLOv5s 网络结构的 Neck 层,为了融合不同尺寸的特征信息,使用了 FPN+PAN 结构。FPN 将高层的特征信息通过上采样的方式进行传递融合,得到进行预测的浅层特征信息,而 PAN 则将浅层特征信息通过下采样的方式,将具有强定位信息传递到高层特征信息,从而实现不同的主干层对不同的检测层进行特征聚合。但 PAN 结构较为简单,缺少原始信息参与学习,易出现训练学习偏差的情况,从而影响检测精度。因此,本文采用 BiFPN[19]加权双向金字塔结构替换 FPN+PAN 结构。图 6 为 BiFPN 结构图。图 6 中,BiFPN 主要作用是加权特征图融合和高效的双向跨尺度连接,其中加权特征融合是在特征融合期间为每个输入添加一个额外的权重,让网络去学习每个特征的重要性。高效的双向跨尺度连接是在 FPN+PAN 结构的基础上,去掉了只有一条输入边和输出边的结点,如果在同一层出现输入结点和输出结点,就添加一条额外的路径,以重复堆叠的方式获得更高级的融合特征。

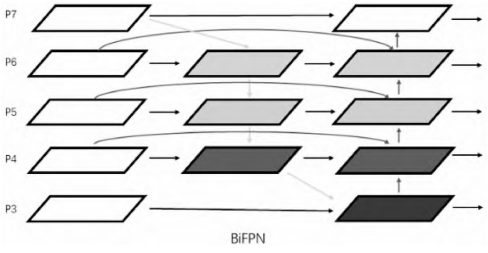

图 6 BiFPN 结构图

3 实验

3.1 实验数据

本文按照 VOC 数据集的格式制作了实验数据集,一共 8462 张行人数据集,其中包含室内 3000 张与室外 5462 张行人数据,如图 7 所示。其中按照 9:1 比例,将训练集和验证集分别划分为 7616 张和 848 张行人数据。

图 7 数据集样例图片

3.2 实验环境

本文在 Ubuntu 18.04 系统下进行模型训练与测试,搭载CPU 为 Intel i7-10700F,2.90 GHz,内存 16 GB,GPU 为 NVIDIA GeForece RTX3090,显存 24 GB。使用 PyTorch 深度学习框架进行模型构建与改进,cuda 版本为 11.1.0,cudnn 版本为 8.1.0,Python 版本为 3.7,batch_size 为 32,epoch 为 300。

3.3 实验结果与分析

3.3.1 模型对比

本次实验通过统一的训练集训练后,在统一的测试集上进行测试。将改进后的模型上分别与 YOLOv3-spp 和 YOLOv5s 进行对比,对比结果如表 1 所示:

表 1 实验结果对比

Models	mAP(%)	Size(MB)	GFLOPs	FPS(frame/s)
YOLOv3-spp	99.4	122.64	155.6	142.86
YOLOv5s	99.4	14.1	15.8	333.33
YOLOv5s-Ghost-ECA-Bifpn	99.3	**6.61**	**6.3**	**555.56**

从表 1 中可以看出在平均精度保持不变的情况下,改进后的模型在模型大小、浮点型计算量都有大幅度减小,并且检测速度都有明显的提升。比 YOLOv3-spp 的模型大小缩小了约 18.56 倍,浮点型计算量减少了约 24.70 倍,检测速度提高了约 3.89 倍;与 YOLOv5s 相比较,模型大小缩小了约 2.13 倍,浮点型计算量减少了约 2.51 倍,检测速度提高了约 1.67 倍。

3.3.2 消融实验

本次实验将不同的模块引入 YOLOv5s 并与 YOLOv5s 进行对比,结果对比如表 2 所示:

表 2 实验结果对比

Models	mAP(%)	Size(MB)	GFLOPs	FPS(frame/s)
YOLOv5s	99.4	14.1	15.8	333.33
YOLOv5s-Ghost	99.2	7.6	8.1	434.78
YOLOv5s-Ghost-ECA	99.3	**6.48**	7.7	434.78
YOLOv5s-Ghost-BiFPN	99.3	7.74	8.3	526.32
YOLOv5s-Ghost-ECA-BiFPN	99.3	6.61	**6.3**	**555.56**

从表 2 中可以看出在给 YOLOv5s 引入不同模块各项指标都出现了不同程度的提升。与 YOLOv5s 原始模型相比较,单独引入 Ghost 卷积模块,模型大小缩小了约 1.86 倍,浮点型计算量减少了约 1.95 倍,检测速度提高了约 1.30 倍;在引入 Ghost 的基础上单独引入 ECA 模块,模型大小缩小了约 2.17 倍,浮点型计算量减少了约 2.05 倍,检测速度与单独引入 Ghost 模块保持一致;在引入 Ghost 的基础上引入 BiFPN 模块,模型大小缩小了约 1.82 倍,浮点型计算减少了约 1.90 倍,检测速度提高了约 1.58 倍。

通过消融实验,可以证明对 YOLOv5s 进行算法结构优化可以有效提高行人检测的速度,减少浮点型计算量以及缩小模型大小。

参考文献

[1]WANG Z Y, ZHANG Y Z, LIU Y, et al.MFC-Net: Multi-feature fusion cross neural network for salient object detection [J].Image and Vision Computing, 2021,113 (Sep.):104243.1-104243.10

[2]GIRSHICK R, DONAHUE J, DARRELL T,et al.Rich fea- ture

(下转第 89 页)

可以看到，图 9 所示的四组融合后的点云图与煤斗内原始煤堆非常接近。

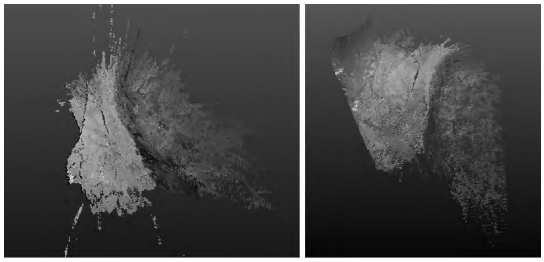

图 9 两组模型融合后点云图(左)及四组模型融合后点云图(右)

然后，我们再将模型融合后的点云图进行网格化，通过五次迭代得到迭代后的网格化图图 10(中)，然后如图 10(右)进行修剪边缘后，将煤斗框套入即得到了完整的煤斗重建图，如图 11 所示。

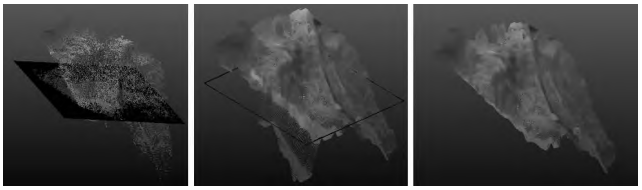

图 10 网格化后点云图(左)、五次迭代后网格化图(中)及修剪边缘后点云图(右)

3 结束语

针对封闭空间内物体和场景的三维重建要求，用两个相同的鱼眼摄像头间隔一定距离安装所形成的摄像机组，来获取双目视觉图。通过对鱼眼双目摄像机组的标定、建模、模型融合等步骤，可以对所拍摄的照片组中物体的轮廓和结构进行精确重建和恢复。本文通过对某热电厂煤斗进行基于鱼眼相机的双目视觉法得到相应的图像对，然后对其进行矫正、特征点匹配、视差图计算、三维重建、噪声滤除以及点云模型融合，最后套上煤斗模型得到最终的重建图，以便开展后续的煤斗内体积检测或者料位检测等工作。本文提供的三维重建模型及方法能较好地恢复煤斗场景，具有一定的应用价值。

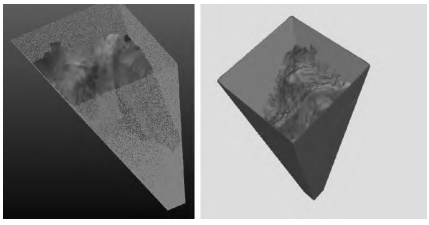

图 11 加入煤斗后重建模型及示意图

参考文献

[1]张阳阳.基于双目鱼眼系统的煤堆体积测量技术研究与应用[D].杭州：浙江大学，2020

[2]郑太雄，黄帅，李永福，等.基于视觉的三维重建关键技术研究综述[J].自动化学报，2020，46(4)：631-652

[3]吴健辉，商橙，张国云，等.鱼眼相机与 PTZ 相机相结合的主从目标监控系统[J].计算机工程与科学，2017，39(3)：540-546

[4]徐建芳，刘志刚，韩志伟，等.基于 SIFT 和 LBP 点云配准的接触网零部件三维重建研究[J].铁道学报，2017，39(10)：76-81

[5]HE H, CHEN T, ZENG H, et al. Ground Control Point-Free Unmanned Aerial Vehicle-Based Photogrammetry for Volume Estimation of Stockpiles Carried on Barges [J]. Sensors: MDPI, 2019, 19(16): 3534-3555

[收稿日期：2022-07-18]

(上接第 86 页)

hierarchies for accurate object detection and semantic segmentation [C] //Proceedings of the IEEE conference on computer vision and pattern recognition.Columbus: IEEE, 2014: 580-587

[3]GIRSHICK R. Fast R-CNN[J]. Computer Science, 2015

[4]Ren SHAOQING, He KAIMING, Girshick Ross, et al. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks [J]. IEEE transactions on pattern analysis and machine intelligence, 2017,39(6):1137-1149

[5]HE KM, GKIOXARI G, DOLLÁR P, et al. Mask R-CNN [C]// Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV). Venice: IEEE, 2017:2980-2988

[6]SONG Y, FU Z. Uncertaain mulivariable regeression model[J]. Soft Computing,2018,22(17):5861-5866

[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: Springer, 2016: 21-37

[8]ZHOU X, WANG D, KRÄHENBÜHL P. Objects as point[J]. ArXivPreprint, ArXiv:1904.07850, 2019

[9]REDMON J, DIVVALA S, GIRSHICK R, et al. You Only Look Once: Unified, Real-Time Object Detection [J]. IEEE conference on Computer Vision and Pattern Recognition, 2016:779-788

[10]REDMON J, FARHADI A. YOLO9000: Better, Faster, Stronger [C]// IEEE Conference on Computer Vision & Pattern Recognition, 2017:6517-6525

[11]Redmon J, Farhadi A. YOLOv3: An Incrementall mprovement [J]. arXiv e-prints, 2018

[12]Bochkovskiy A, Wang CY, Liao H. YOLOv4: optimal speed and accuracy of object detection[J]. arXiv: 2004. 10934

[13]Lin T Y, Dollár P, Girshick R, et al. Feature pyramid networks for object detection [C] //The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017: 5936-5944

[14]Liu S, Qi L, Qin H F, et al. Path aggregation network for instance segmentation [C] //2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018

[15]Han K, Wang YH, Tian Q, et al. GhostNet: More features from cheap operations [C] //Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020:1577-1586

[16]赵文清，杨盼盼.双向特征融合与注意力机制结合的目标检测[J].智能系统学报，2021，16(6)：1098-1105

[17]Jie Hu, Li Shen, Gang Sun. Squeeze-and excitation networks[J].arXiv preprint arXiv:1709.01507 7, 2017

[18]WANG Q L, WU B G, ZHU P F, et al. ECA-Net: efficient channel attention for deep convolutional neural networks[C] //2020 IEEE/ CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020: 11531-11539

[19]TAN M,PANG R, LE Q V.EfficientDet:Scalable and efficient object detection[C] //Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020:10781-10790

[收稿日期：2022-10-13]