并行多尺度特征增强的小目标检测

侯晓辉[1]，李 莉[2]，孙红凯[1]，马 亮[1]

（1. 中国大唐集团有限公司，内蒙古 赤峰 024000；2. 华北电力大学 新能源学院，北京 102206）

摘要：由于小目标像素点少，本身携带的特征较少，大多数目标检测算法不能有效利用特征图中小目标的边缘信息和语义信息，导致小目标检测精度低，漏检、误检现象时有发生。为解决 RetinaNet 模型小目标信息特征不足的缺陷，在 RetinaNet 模型中引入一个并行辅助的多尺度特征增强模块 MFEM(muti-scale feature enhancement model)，通过使用不同扩张率的空洞卷积，避免多次下采样造成信息损失的同时，有利于辅助浅层提取多尺度上下文信息。另外，采用一种专门针对目标检测任务而设计的主干网改进方案，可有效保存高层特征图的小目标信息。传统自上而下的金字塔结构侧重于将高层语义从顶层传递到底层，单向传递的信息流不利于小目标的检测。将辅助 MFEM 分支与 RetinaNet 相结合，构造一个包含双向特征金字塔结构的模型，它可有效地融合网络高层强语义信息和底层高分辨率信息。为证明文中算法 FE-RetinaNet (Feature Enhancement RetinaNet)的有效性，在 MS COCO 公共数据集进行实验。与原 RetinaNet 相比，改进的 RetinaNet 在 MS COCO 数据集上检测精度(mAP)取得了 1.8%的提升，COCO AP 为 36.2%；FE-RetinaNet 在小目标上检测效果良好，APs 提高了 3.2%。

关键词：目标检测；小目标；特征增强

中图分类号：TP312 **文献标志码**：A **文章编号**：1671－1807(2023)05－0178－11

目标检测任务一直是计算机视觉领域的主要任务之一。近年来，得益于卷积神经网络的发展，目标检测算法的性能取得了长足进步。基于卷积神经网络的目标检测模型可分为两类：一类是两阶段目标检测模型，如 R-CNN[1]、Fast R-CNN[2]、Faster R-CNN[3]、Mask R-CNN[4]、R-FCN[5] 等。两阶段的目标检测算法首先在第一阶段区分前景和背景，产生候选区域，然后在第二阶段提取特征，对目标进行分类和位置回归。另一类是单阶段目标检测模型，如 SSD[6]、YOLO[7]、YOLO9000[8]、YOLOv3[9]、YOLOv4[10]、RetinaNet[11]、DANet[12] 等。单阶段目标检测算法则直接产生物体的类别概率和位置坐标值，经过单次检测即可直接得到最终的检测结果。总的来说，相对于两阶段目标检测模型，单阶段检测模型速度更快，但对目标（尤其是小目标）的检测精度却稍逊一筹。

近年来，研究人员一直致力于在保证单阶段目标检测模型速度的同时，尽可能提升模型的精度，从而涌现了非常多速度快、精度高的单阶段目标检测算法。文献[13]构造了一个自上而下的特征金字塔结构，通过融合高层强语义特征图和低层高分辨率特征图，使高层语义信息传递到低层，通过融合多尺度信息提高网络的精度。文献[11]设计了一个损失函数 focal loss，解决了单阶段目标检测中正负样本比例严重失衡的问题，使得 RetinaNet 精度比当时最先进的两阶段检测模型还高。在现有的目标检测算法中，RetinaNet 通过引入 focal loss 解决了类别不平衡问题，得到了研究人员的广泛关注。标准的 RetinaNet 框架使用 ResNet 为主干网，通过自顶向下的特征金字塔结构融合多尺度语义信息，并在网络的后端加入两个子网分别用于目标分类和边界框回归。然而标准的 RetinaNet 对于大、中目标检测效果良好，对于小目标的检测效果却不尽如人意。

通过对 RetinaNet 的分析与研究，本文找到了两个对 RetinaNet 检测小目标性能影响最为关键的

收稿日期：2022-10-26

作者简介：侯晓辉(1974—)，男，内蒙古赤峰人，中国大唐集团有限公司，赤峰新能源事业部总经理，高级经济师，高级工程师，研究方向为机器学习、新能源、智能风电场；李莉(1974—)，女，山东济南人，华北电力大学新能源学院，副教授，博士，研究方向为风能资源评估、风电功率预测、风电场流场数值模拟；孙红凯(1986—)，男，内蒙古承德人，中国大唐集团有限公司，赤峰新能源事业部生产部主任，工程师，研究方向为新能源、智慧风电场；马亮(1986—)，男，内蒙古赤峰人，中国大唐集团有限公司，赤峰新能源事业部生产部专工，研究方向为新能源、智慧风电场。

因素。首先,小目标信息在特征提取网络的深层和浅层都会遭到不同程度的丢失。一方面,传统的特征提取网络(如 VGG16[14] 和 ResNet[15]),都是重复执行卷积和最大池化下采样操作来提取特征。虽然保留了一定程度的语义信息,但浅层很难提取到多尺度的上下文信息来区分背景和目标。本文使用一个辅助的多尺度特征增强模块,辅助提取多尺度浅层特征,并与主干网提取的特征相融合,大大提高了小目标的表达能力。另一方面,目标检测算法通常使用专门为图像分类而设计的主干网,导致小目标在网络高层特征图中被忽略。较大倍数的下采样特征图虽然具有高级语义信息,但是高层特征图分辨率大大减小,不利于目标的定位。这一问题在 RetinaNet 同样存在:RetinaNet 使用 ResNet 为特征提取网络,额外附加了 P_6 和 P_7 两个特征层来提高大目标的检测效果,导致 P_7 的分辨率大小减少到原来的 1/128,使小目标在该层特征图中不可见。即使 FPN 结构将深层的强语义信息与浅层融合,但由于小目标在如此深层的特征图中缺失,小目标的语义信息也随之丢失。目标检测任务不同于分类任务,不仅需要判断物体的类别,还需要定位目标在空间上的位置。本文旨在不降低大目标检测效果的同时,提高小目标的检测效果。受文献[16-17]的启发,本文设计的主干网可以在不增加下采样次数的前提下,使深层特征图也能保留较大的感受野,有利于保留深层特征图中的小目标信息。

其次,RetinaNet 采用传统自上而下 FPN 结构,使得特征传递会受到单向信息流的限制。虽然 RetinaNet 使用的金字塔结构可以将来自顶层的强语义信息传递到底层,但这种结构只能向前一层注入高级语义信息,而忽略了浅层高分辨率特征图的传递。另外,由于浅层的空间特征信息在自下向上的传递过程中逐步被淡化,使得小目标在层层卷积下采样后丢失了高分辨率特征信息的指导,导致模型检测性能下降。文献[18-19]等虽然都采用了双向特征金字塔结构来实现多尺度特征融合,添加额外的分支缩短浅层特征传播的过程,但这些工作都只是在额外分支中重用了主干网提取的浅层特征,无法捕获丰富的上下文信息。本文则是将多尺度特征增强模块与双向特征金字塔结构相结合,极大地丰富了浅层多尺度上下文信息。

总而言之,通过对 RetinaNet 模型的研究与分析,本文提出了一种具有以下贡献的单阶段目标检测算法,以提高对小目标的检测精度:

1)提出一个简单有效的并行多尺度特征增强模块,利用空洞卷积不进行下采样操作也可以拓展感受野的特点,辅助主干网提取具有多尺度上下文信息的浅层特征。

2)引入一种专门为目标检测任务而改进主干网的方法,有效减少了特征提取网络在检测任务和分类任务之间的差距,使主干网高层特征图在保留大感受野、强语义信息的同时,尽可能保存小目标的纹理信息。

3)将辅助的多尺度特征增强模块与原 FPN 结构结合,构造了一个包含多尺度浅层信息的双向特征金字塔结构。与大多数的双向结构在额外的分支重用主干网提取特征的方法不同,本文则是将多尺度特征增强模块作为额外分支的输入,为网络带来了全新的特征信息。

1 相关工作

小目标检测任务一直是计算机视觉领域的一个挑战。对于大多数单阶段目标检测模型来说,包括 SSD[6]、YOLO[7-10]、RetinaNet[11] 等,都在小目标检测任务上表现不佳。回顾目标检测算法的发展,提高目标检测精度的方法主要是采用更好的特征提取网络、添加更多的上下文信息和多尺度特征融合。

1)更好的特征提取网络。大多数目标检测算法通过采用更好的特征提取网络来提高检测算法的精度。SSD[6] 使用 VGG-16 为基本网络,提取特征的性能相对较弱,这也是 SSD 检测效果相对较差的原因之一。总而言之,先进的目标检测算法大都采用更深的网络来提取特征。由于较大物体在比较深的特征图上进行预测,对应原图比例的感受野需求也大,所以 RetinaNet 又在 ResNet 的基础上额外添加了 P_6 和 P_7 两层,来提高对大目标的检测效果。但一方面,特征图越深,物体边缘清晰度越模糊,对应的回归就较弱;另一方面,深层特征图分辨率小,小物体在深层特征图上便很难可见。即使 FPN[13] 及 RetinaNet[11] 等网络把浅层特征与语义强的深层特征相加,由于小物体目标在深层特征图中已经消失,很大部分的语义信息还是未得到保留。文献[16]为解决该问题,构造了一个专用于检测任务的网络,在主干网中采用空洞卷积降低了下采样次数,保持大感受野的同时也保留了小目标的纹理信息。络用于提取浅层特征。然而,模型从零开始训练可能会遇到收敛问题,导致辅助网络的性能甚至比预训练的更差。而本文使用的多尺度特征增强模块与之不同,它在网络中作为一个辅助分

支可以进行端对端训练。

更多的上下文信息：有许多工作致力于增加边界框之外的信息，即添加更多的上下文特征来提高目标检测的精度。文献[20]在特征融合模块中添加自适应路径融合网络来融合更多的位置信息和语义信息，提高了多尺度目标的检测精度。文献[21]提出要增加浅层网络的层数以及采用*K*-means聚类算法选取初始先验框，提高了网络像素特征提取细粒度并加快了检测速度。Inception系列设计了一系列多分支的卷积结构，通过增加神经网络的宽度来提升训练效果。该结构在每个分支分别设置不同大小的卷积核，通过提取到感受野不同大小的特征，来丰富上下文信息。但由于Inception人工设计的痕迹过重，ResNex则将Inception结构与残差结构结合，加入了多分组卷积的思想，构造了一个多分支模型结构。该模块每个分支都有相同的拓扑结构，使模型实现了高度模块化。RFBNet则使用空洞卷积生成更大的感受野，在更大的区域捕捉到更多的上下文信息，同时使特征图保持较高的分辨率。本文的辅助多尺度特征增强模块便是受此类结构的启发，在多分支获取不同大小的感受野，为浅层添加更多的上下文信息。

多尺度特征融合：FPN通过自顶向下和横向连接的结构，将低分辨率、强语义特征图与高分辨率、低语义特征图结合起来，用于在所有尺度上构建高级语义特征映射，这对小目标检测很重要。遵循这一思想，为了缩短底层和顶层特征之间的传递路径，文献[17]在FPN的基础上添加了自底向上路径增强支路，进一步增强了网络的定位能力；文献[12]也引入了一个加权的双向特征金字塔结构，有利于网络进行简单快速的多尺度特征融合。然而，这些双向特征金字塔结构大多是复用主干网提取的特征，而本文在此基础上添加了额外的特征增强模块，提高了信息的多样性及丰富性。

2　方法

本节首先介绍设计的检测模型的总体结构，然后分别介绍3个主要部分，包括改进的ResNet-D、并行辅助多尺度特征增强模块MFEM(muti-scale feature enhancement model)及双向特征金字塔结构(Bidirectional feature pyramid structure)。本文的双向特征金字塔结构同时利用了ResNet-D提取的特征和多尺度特征增强模块提取的多尺度上下文特征，在加快浅层特征传递效率的同时提高了网络对小目标的表达能力。

2.1　总体结构

Wu等[22]考虑到行人重识别在实际场景的应用问题，提出跨模态行人重识别这一研究问题，并公开了数据集SYSU-MM01。数据集分别由6部相机采集，其中1、2、4、5号相机拍摄RGB图像，3、6号相机拍摄红外图像，并且1、2、3是在室内环境。

图1显示了网络的整体结构，它主要由3个部分组成：改进的ResNet、多尺度特征增强模块MFEM及双向特征金字塔结构。受DetNet启发，本文改进了ResNet并命名为ResNet-D，该主干网抛弃了原RetinaNet中64及128倍的下采样，使用膨胀残差结构代替常规的残差模块，使高层特征图具有较大的感受野的同时，保留更多小目标的纹理信息。

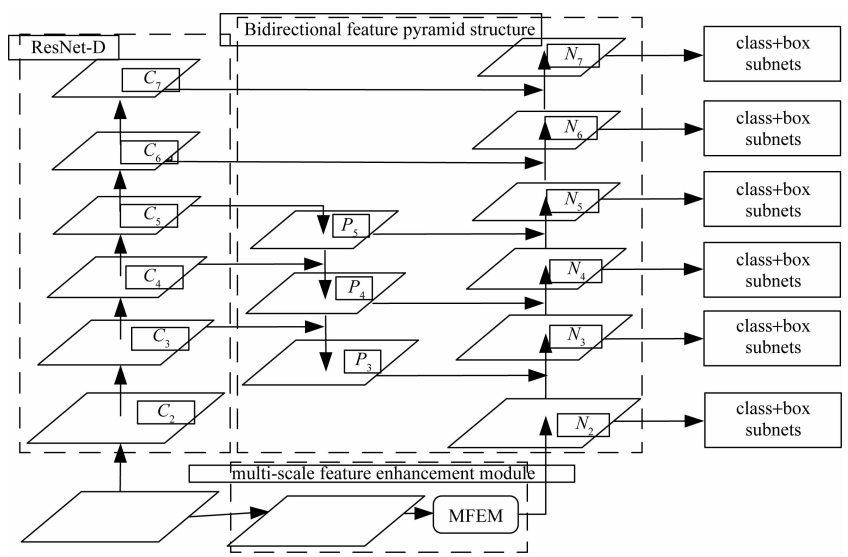


图1　本文方法整体网络结构

RetinaNet 使用固定感受野大小的卷积核提取特征,忽略了目标上下文信息对小目标检测的影响。受 Inception 结构和 ResNext 分组卷积思想的启发,本文设计了多尺度特征增强模块 MFEM,MFEM 利用 4 个具有不同膨胀率大小的膨胀卷积分支,提取出具有不同感受野的特征图,即提取更多目标之外的特征以提高标准 RetinaNet 预测层的鉴别能力。

浅层上下文信息:浅层特征有更好的分辨率,小目标往往更依赖浅层。具体来说,由于浅层特征图分辨率大,有利于小目标检测,但它是一个具有不同感受野的多分支结构,可以极大丰富浅层上下文信息。将多尺度特征增强模块与 RetinaNet 结合,构造了一个融合特征增强模块的双向特征金字塔结构。该结构可以将特征增强模块提取的浅层高分辨率特征通过自底向上的分支融合,有效提高小目标的检测精度。

2.2 改进的主干网 ResNet-D

本文从数据集中随机选取了一部分分析该数据集的特点,上面一行是 RGB 图像,下面一行是 IR 图像。纵向来看,两种图像的存在巨大的模态差异,RGB 图像具有丰富的颜色信息,外观特征非常明显,而红外图像仅仅具有一些纹理特征和轮廓特征,二者存在严重的特征信息不匹配。横向来看,模态内图像之间姿态、视角变化很大,部分图像还存在遮挡问题,进一步给特征匹配增加了困难。综上所述,SYSU-MM01 数据集具有很大的挑战性,需要从模态差异与同身份差异两个角度出发综合解决问题.

现有检测框架中常用的特征提取策略通常是重复堆叠多个卷积层和最大池化层(如 ResNet-50),来构造更深的特征提取网络,以产生强语义的信息。这样的特征提取策略对更倾向于平移不变性的图像分类任务更有利。与图像分类不同的是,目标检测还需要精确的对象描述,而局部低、中层次特征(例如纹理)信息也是关键。

主干网的设计通常存在两大难题:①使主干网保持较高的空间分辨率会极大地消耗内存和时间;②减少下采样次数会导致感受野减小,不利于大目标检测。受 DetNet 的启发,本文改进了 ResNet,旨在不降低大目标检测精度的同时,提高小目标的检测效果。本文引入了一个新的残差结构,即将标准残差模块中的 3×3 标准卷积层替换为卷积核大小为 3×3,膨胀系数为 2 的空洞卷积,形成了膨胀残差结构,可有效扩大特征图的感受野,如图 2(b)所示。另外,本文在膨胀残差结构的 shortcut 分支加入了 1×1 卷积层,使第 7 阶段不经过下采样也能获得全新的语义信息,如图 2(c)所示。

标准的 RetinaNet 框架采用 ResNet 为主干网提取特征[图 3(a)],并额外添加了两层特征 P_6 和 P_7(P_6 是在 C_5 上进行卷积核为 3×3,步长为 2 的卷积运算得到的,P_7 是通过在 P_6 上应用 ReLU 以及卷积核为 3×3,步长为 2 的卷积运算得到的),虽有利于大目标的检测,但会使得高层特征图由于下采样次数过多,导致小目标被忽略[图 3(b)]。为使高层特征图具有较大的感受野的同时,保留更多小目标的纹理信息,本文在主干网络上引入了更多的阶段(即 C_6 和 C_7),在 stage6-7 中将常规的残差模块替换为加入膨胀卷积的膨胀残差结构[图 3(c)]。

改进的 ResNet-D 网络结构如图 4 所示,本文在 ResNet 原 5 阶段的基础上额外添加了两个阶段——第 6 及第 7 阶段(stage 6 及 stage 7),旨在遵循 RetinaNet 设计思想,来提升网络对大目标的检测效果。ResNet-D 保持前 5 个阶段均使用原残差结构,第 6 和第 7 阶段使用由 3 个残差模块组成的膨胀残差结构,可将两个阶段的空间分辨率固定为 32 倍下采样,有效扩大高层特征图的感受野。修改后网络每阶段的下采样倍数分别为[2,4,8,16,32,

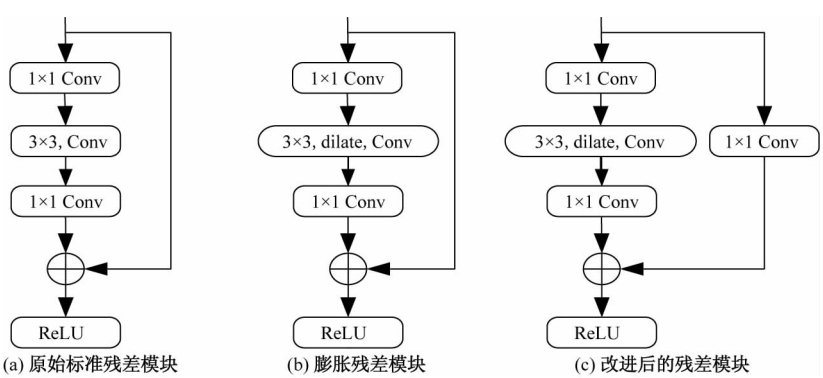


图 2 膨胀残差结构

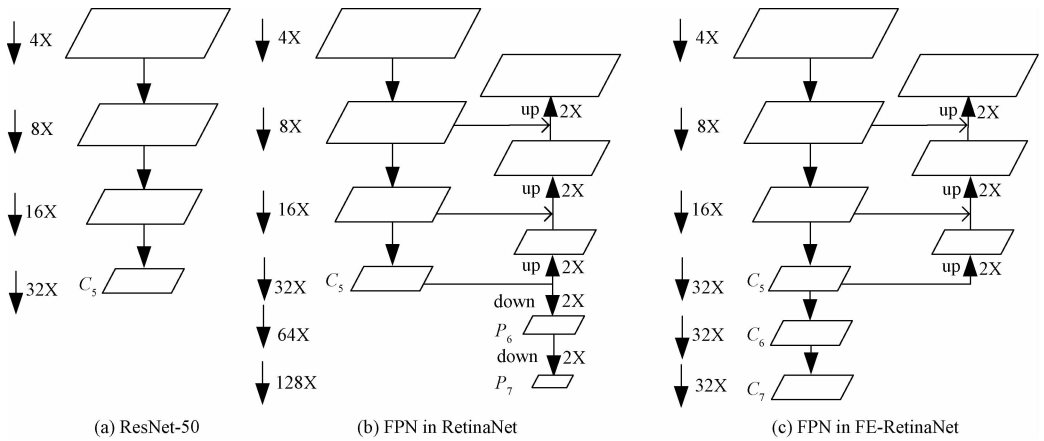


图 3　网络结构对比图

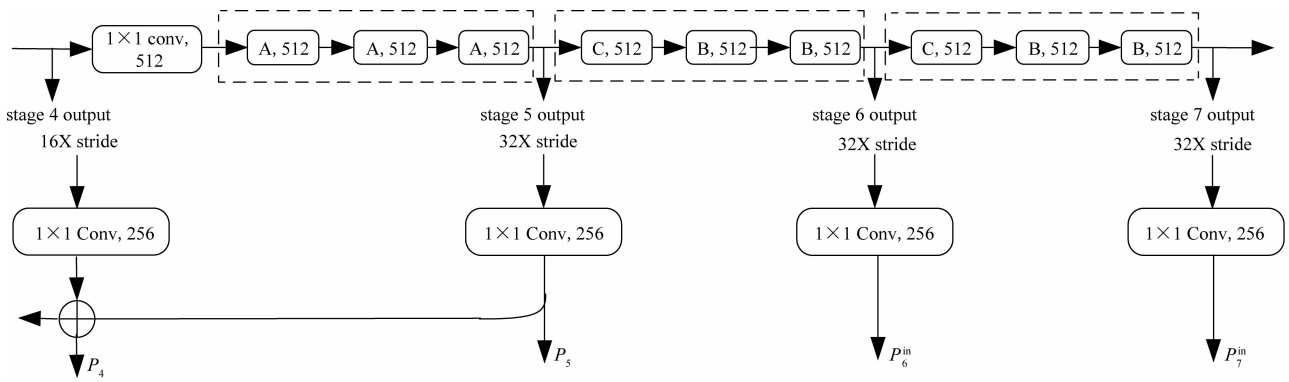


图 4　ResNet-D 网络图

32,32],抛弃了 RetinaNet 中 64 倍和 128 倍下采样的特征图,而采用膨胀瓶颈结构来保持小目标纹理信息,并扩大特征图的感受野。第 6 和第 7 阶段的每层卷积有 3 个膨胀瓶颈结构,以(C,B, B)的顺序进行排列。由于该阶段使用了膨胀卷积而非常规的卷积结构,所以有效避免了下采样操作。此外为减少网络的计算量,ResNet-D在第 4 阶段后使用 1×1 卷积将特征图降维到 512 通道,一直到第 7 阶段特征图通道数均为 512。ResNet-D 保持前 5 个阶段均使用原残差结构,该阶段依然进行下采样操作。其中第 4 到第 5 阶段采用原结构作为卷积层,该阶段依然进行下采样操作。第 5 至第 6 以及第 6 至第 7 阶段为保持 32 被下采样,将上文设计的膨胀瓶颈结构进行组合。

2.3　多尺度特征增强模块

本文与文献[10]中的实验方案一致，将红外图像作为 query 集，RGB 图像作为 gallery 集，从 single-all、single-indoor、multi-indoor、multi-all 4 种模式分别测试方法的有效性，同样以标准累计匹配特性(CMC)曲线和平均精度(mAP)作为性能评价指标。由于大多数检测算法采用固定尺寸的卷积核提取目标特征,只能提取局部的特征信息,感受野大小受到了限制,无法捕获丰富的上下文信息,不利于检测复杂的自然图像。本文受 Inception 获取不同大小感受野和 ResNext 分组卷积思想的启发,设计了具有不同感受野的多分支结构,先采用不同扩张率的空洞卷积获取不同尺度的信息,再融合不同尺度的信息获取丰富的上下文信息。多尺度特征增强模块(MFEM)是一个简单的结构,它是一个辅助的分支直接与 ResNet-D 连接,用于辅助提取携带多尺度上下文信息的浅层特征。本节介绍了在 MFEM 中使用的特征提取策略,然后描述了 MFEM 体系结构。

多尺度特征增强模块 MFEM:标准的 RetinaNet 框架使用 ResNet 提取特征,特征提取都是由卷积和最大池化重复执行的。虽然下采样过程保留了一定程度的语义特征,但是仍有可能会失去有助于检测的低级特征。另外,当前的特征提取网络通常将同一阶段的感受野设置为相同的大小,这会

导致特征在判别性和鲁棒性方面的损失。为了弥补浅层特征在下采样过程中的丢失,提高特征的判别性和鲁棒性,本文的 MFEM 模块提供了一种辅助的特征提取方案,如图 5 所示。

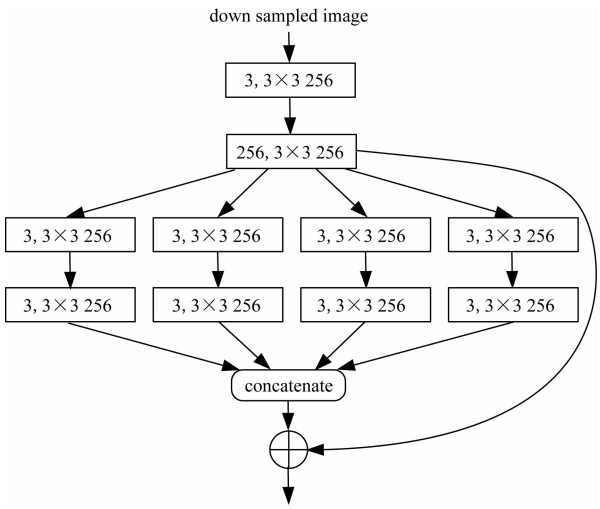


图 5　多尺度特征增强模块结构图

首先,通过简单的池化操作对输入图像 I 进行 4 倍下采样得到 I_t,使其大小与 ResNet-D 的第 2 阶段特征图 C_2 相匹配。由此产生的下采样图像(down sampled image)通过模块化的分组卷积结构,包括分割、转换、聚合操作,输出包含多尺度上下文信息的特征图。

下采样图像 I_t 首先通过大小为 3×3 和 1×1 的两个连续卷积层,产生初始特征映射 $F_{\text{int}(0)}$,即

$$F_{\text{int}(0)} = \varphi_{\text{int}(0)}(I_t) \tag{1}$$

式中:$\varphi_{\text{int}(0)}$ 为一个串行操作,包括一个 3×3 卷积和一个 1×1 卷积的模块。然后利用初始特征映射 $F_{\text{int}(0)}$ 生成中间特征集 $F_{\text{int}(k)}$,即

$$F_{\text{int}(k)} = \omega_{\text{int}(k)}(F_{\text{int}(0)}) \tag{2}$$

式中:k 为多尺度特征增强模块的分支数,本文中共有 4 个分组卷积结构;$\omega_{\text{int}(k)}$ 为分组卷积的第 k 个分组,包括一个 1×1 卷积层和一个卷积核大小为 3×3,膨胀率分别为 1,2,3,4 的膨胀卷积层。然后将 4 个包含多尺度上下文信息的特征图进行 concatenate 方式特征融合,得到融合后的特征 F_{concat},即

$$F_{\text{concat}} = \sum_{k=1}^{4} \text{Concat}(F_{\text{int}(k)}) \tag{3}$$

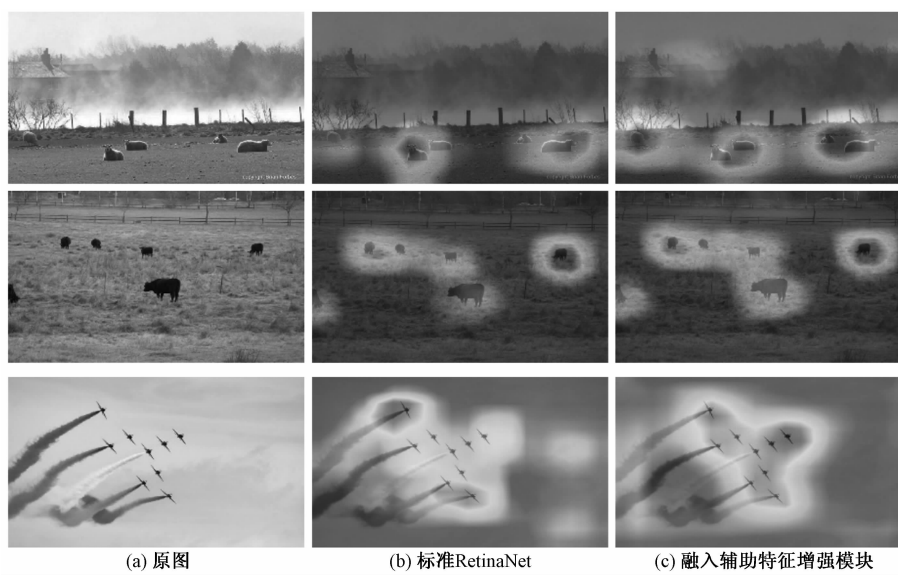
式中:Concat 为通道间信息的合并。随后将初始特征映射 $F_{\text{int}(0)}$ 通过 shortcut 支路进行恒等映射,特征 F_{concat} 进行 addition 方式特征融合,得到融合后的特征为 F_{add},即

$$F_{\text{add}} = F_{\text{concat}} \oplus F_{\text{int}(0)} \tag{4}$$

式中:⊕表示 addition 方式特征融合。将 F_{add} 通过 1×1卷积层进行升维,得到输出特征 N_2,以备将这些转换后的特征通过自下而上的金字塔分支进行聚合:

$$N_2 = \text{Conv}(F_{\text{add}}) \tag{5}$$

如图 6 展示了融入多尺度特征增强模块后的特征热力图对比。对于 MS COCO 数据集的原始图像,标准 RetinaNet 的输出特征为图 6(b)所示,可以看出 RetinaNet 对小目标不敏感;融入辅助特征增强模块后,如图 6(c)列所示,可发现本文网络的特征热力图可以更好地覆盖物体边界,这证明了辅助多尺度特征增强模块可以有效丰富利用小目标检测的特征,使网络更能关注到被忽略的小目标。



(a) 原图　(b) 标准RetinaNet　(c) 融入辅助特征增强模块

图 6　多尺度特征增强模块后的特征热力图对比

2.4 双向特征金字塔结构

如前所述，浅层特征和深层特征对于小目标的检测都很重要，然而 RetinaNet 采用的自上而下的单向金字塔结构侧重于将强语义信息从顶层传递到底层，忽略了低层高分辨率特征的传递，它具有丰富的定位信息。另外，标准的 RetinaNet 框架使用 ResNet 提取特征，重复执行下采样操作会造成低层特征丢失。双向特征融合旨在融合高低层不同语义信息、不同分辨率的特征，同时加快浅层特征传播的效率。

自顶向下的分支：本文遵循 FPN 的定义，将产生相同空间大小的特征层定义为处于同一网络阶段，每个特征级别对应一个网络阶段。将 2.2 小节改进的 ResNet-D 作为基本结构，选用$\{C_3,C_4,C_5\}$作为特征层。自顶向下的路径将高层语义信息更强的特征，通过横向连接自上而下进行融合。与标准 RetinaNet 一致，本文对每一个低分辨率特征图进行上采样处理，将空间分辨率扩大 2 倍，来和下一层特征图大小相匹配。自顶向下的路径使用第 3 至第 5 阶段特征作为输入，即

$$\boldsymbol{C}=(C_3,C_4,C_5) \tag{6}$$

式中：C_i 为分辨率为原图 $1/2^i$ 大小的输出特征图。例如，对于一个输入为 512×512 大小的图像，那么 C_3 的分辨率大小为 64×64。对于每个横向连接路径，本文通过一个 1×1 卷积层将每个特征图降维到 256 通道，得到$\{P_3^{\text{in}},P_4^{\text{in}},P_5^{\text{in}}\}$，分别对应于$\{C_3,C_4,C_5\}$，表示为

$$P_i^{\text{in}}=\text{Conv}(C_i),\ i=3,4,5 \tag{7}$$

式中：Conv 为 1×1 卷积运算，用于将每一个阶段的输出特征降维到 256 通道。

本文采用自顶向下、横向连接的方法构造自上向下的通路。自顶向下的路径可以产生分辨率更高、语义特征丰富的特征，这些特征通过横向连接的路径自顶向下增强。每个横向连接将具有相同大小的特征映射融合为一个阶段。自顶向下特征融合过程可表示为

$$P_i=P_i^{\text{in}}\oplus\text{ReSize}(P_{i+1}),\ i=3,4 \tag{8}$$

$$P_5=P_5^{\text{in}} \tag{9}$$

式中：⊕表示特征融合操作；ReSize 为上采样操作，目的是与下层要融合的特征图分辨率相匹配。

自底向上的分支：本文没有像 PANet 一样，在自底向上的分支中复用主干网提取的特征，而是创新性地将 2.3 节描述的多尺度特征增强模块作为自底向上分支的输入。由于辅助特征增强模块的每个分支具有不同的感受野，那么将经过该结构的特征作为输入，可以丰富多尺度浅层上下文信息。

本文将经过特征增强模块的输出特征 N_2 作为第一提取层，它与 C_2 具有相同空间分辨率大小，随后的$\{N_3,N_4,N_5\}$的空间分辨率分别与$\{P_3,P_4,P_5\}$相对应。最后将 C_6 和 C_7 降维到 256 通道，分别与 N_6、N_7 进行特征融合，该阶段无下采样过程。与自顶向下的分支类似，自底向上分支仍然采用横向连接的方式，将$\{P_3,P_4,P_5,C_6,C_7\}$作为输入，得到输出特征N_i：

$$P_i^{\text{in}}=\text{Conv}(C_i),\ i=6,7 \tag{10}$$

$$N_i=\text{Resize}(N_{i-1})\oplus P_i,\ i=3,4,5 \tag{11}$$

$$N_i=N_{i-1}\oplus P_i^{\text{in}},\ i=6,7 \tag{12}$$

式中：⊕表示特征融合操作；Conv 表示 1×1 卷积运算，用于将该阶段的输出特征降维到 256 通道；ReSize表示下采样操作，目的是与上层特征图的分辨率相匹配。

3 实验及结果分析

为证明 FE-RetinaNet 的有效性，本文在 MS COCO 2017 上数据集进行了实验。本文实验条件配置如下：CPU 为 Intel i7-9700 k；内存为 32 G；GPU 为 NCIDIA GeForce GTX TITAN X；深度学习框架为 Pytorch 1.7.1；CUDA 版本为 10.1。

3.1 数据集和评估指标

MS COCO 数据集是一个由 80 个不同对象类别组成的庞大数据集，具有大量的小目标（占所有目标对象的 41%），特别适合评估小目标的检测效果。按照通用的划分标准，本文将 MS COCO trainval35k 作为训练集（其中包括 80k train 训练集和 35k val 验证集子集），将 minival 作为验证集（由剩余的 5k 验证集组成），将 test-dev 作为测试集来评估本文的模型。本文模型训练时均采用批量随机梯度下降法（stochastic gradient descent，SGD）来优化损失函数，动量参数设置为 0.9，batchsize 设置为 32，初始学习率设置为 0.001。对于第一次 160 k 迭代，本文使用10^{-3}的学习速率，然后对于 60 k 迭代使用10^{-4}，对于最后 20 k 使用10^{-5}。

在 MS COCO 数据集的实验中，使用的评价指标是 mAP。对于 IoU 采用步长为 0.05，0.5～0.95 的设定。对于不同的 IoU，计算对应的 mAP，最后计算所有 mAP 的均值。对于像素区域大小小于 32×32的物体记为小目标，大于 32×32 并且小于

96×96 的物体记为中等目标,大于 96×96 的物体记为大目标。最后统计不同大小物体相应的准确率:小目标准确率 AP_s、中等目标准确率 AP_m 和大目标准确率 AP_l。

3.2 MS COCO 数据集

表 1 展示了本文提出的算法与其他算法在 MS COCO test-dev 数据集上的结果比较。大多数的两阶段检测算法依赖于更大的输入图像来提高模型的性能。在单阶段检测算法中,本文取～500×500 的输入进行比较。对于 832×500 的输入,标准 RetinaNet 获得了 AP 为 34.4%的实验结果。对于大目标的检测结果(AP_l)来说,RetinaNet 获得了 AP 为 49.1%的优秀结果,然而它在小目标的检测精度(AP_s)却不尽如人意,APS 为 14.7%。本文算法以改进后的 ResNet-101 为主干网,在各项指标中均超越了以 ResNet-101 为主干网的 RetinaNet,获得了 1.8%的 AP 增益。更重要的是,在小目标的识别中本文算法取得了更加良好的效果,APS 提高了 3.2%。

表 1 MS COCO 数据集检测结果

算法		Backbone	输入	AP	AP_{50}	AP_{75}	AP_s	AP_m	AP_l
两阶段检测算法	Faster R-CNN	VGG-16	～1 000×600	24.2	45.3	23.5	7.7	26.4	37.1
	Faster R-CNN w FPN	ResNet-101	～1 000×600	36.2	59.1	39.0	18.2	39.0	48.2
	Cascade R-CNN	ResNet-101	～1 280×800	42.8	62.1	46.3	23.7	45.5	55.2
	CoupleNet	ResNet-101	～1 280×800	34.4	54.8	37.2	13.4	38.1	50.8
	R-FCN	ResNet-101	～1 000×600	29.9	51.9	—	10.8	32.8	45.0
	Mask R-CNN	ResNet-101	～1 280×800	38.2	60.3	41.7	20.1	41.1	50.2
单阶段检测算法	YOLOv2	DarkNet-19	544×544	21.6	44.0	19.2	5.0	22.4	35.5
	SSD513	ResNet-101	513×513	31.2	50.4	33.3	10.2	34.5	49.8
	RetinaNet	ResNet-50	～832×500	32.5	50.9	34.8	13.9	35.8	46.7
	RetinaNet	ResNet-101	～832×500	34.4	53.1	36.8	14.7	38.5	49.1
	YOLOv3	DarkNet-53	608×608	33.0	57.9	34.4	18.3	35.4	51.1
	RefineDet	VGG-16	512×512	33.0	54.5	35.5	16.3	36.3	44.3
	DSSD513	ResNet-101	513×513	33.2	53.3	35.2	13.0	35.4	51.1
	RFBNet	VGG-16	512×512	33.8	54.2	35.9	16.2	37.1	47.4
	RFBNet-E	VGG-16	512×512	34.4	55.7	36.4	17.6	37.0	47.6
	EfficientDet-D0	EfficientNet	512×512	34.6	53.0	37.1	—	—	—
	EFIP	VGG-16	512×512	34.6	55.8	36.8	18.3	38.2	47.1
本文算法	FE-RetinaNet	ResNet-50	512×512	34.2	52.8	37.1	16.8	37.2	47.6
	FE-RetinaNet	ResNet-101	512×512	36.2	56.4	39.3	18.0	39.7	49.9

注:～表示输入接近此大小,此网格模型为动态尺寸输入。

3.3 消融实验

为了验证 FE-RetinaNet 中提出策略的有效性,本文研究了所提出改进(ResNet-D、双向特征金字塔结构和多尺度特征增强模块 MFEM)对检测性能的影响,以 ResNet-50 为主干网在 COCO minival 数据集上进行了消融研究,结果汇总见表 2。

可以看出,本文提出的主干网改进策略及特征增强模块均可有效提高算法的检测性能,AP 分别增长了 0.5%和 0.6%。值得注意的是,本文算法提出的改进在小目标的检测取得了较大增益。使用 ResNet-D 为主干网后,AP_s 增长了 0.7%;融入多尺度特征增强模块 MFEM 使 AP_s 增长了 1.6%。此外,通过融入双向特征金字塔结构将算法的总体性能从 32.8%提升到了 33.5%,该实验中自上而下的分支重用了主干网特征,而没有添加 MFEM 结构。

为了进一步证明辅助多尺度特征增强模块的有效性,本文依次添加不同膨胀率的分支,在 COCO minival 数据集上进行实验,结果见表 3。

表 2 融入不同模块算法效果对比

RetinaNet	ResNet-D	BidirectionalFPN	MFEM	AP	AP_{50}	AP_{75}	AP_s	AP_m	AP_l
√				32.3	50.6	34.5	13.7	35.5	46.3
√	√			32.8	51.3	35.2	14.4	35.9	46.8
√	√	√		33.5	51.9	36.1	15.1	36.1	47.1
√	√	√	√	34.1	52.7	36.9	16.7	37.0	47.5

不同膨胀率的分支使小目标的检测结果(AP_s)分别获得了 0.5%、0.6%和 0.4%的精度提升。在 MFEM 中同时添加 4 个不同膨胀率的分支时，本文模型得到了 AP 为 34.2%的最优结果，在小目标上的精度 AP_s 为 16.8%，这表明了本文设计的 MFEM 结构可通过获取浅层上下文特征来提高小目标的检测效果，辅助多尺度特征增强模块对于小目标检测有明显的性能提升。

表 3　添加不同膨胀率分支效果对比

FE-RetinaNet	r1=1	r2=2	r3=3	r4=4	AP	AP_s	AP_m	AP_l
√	√				33.6	15.3	36.2	47.0
√	√	√			33.8	15.8	36.5	47.2
√	√	√	√		34.1	16.4	36.8	47.3
√	√	√	√	√	34.2	16.8	37.0	47.5

3.4　公共数据集检测效果主观评价

为了更加直观地表明本文提出改进对检测性能的提升，在 MS COCO 数据集选取若干图片，分别使用 FE-RetinaNet 与标准 RetinaNet 进行检测，结果如图 7 所示。图 7(a)列为原图，图 7(b)列为标准 RetinaNet 检测结果，图 7(c)列为本文改进的方法检测结果。可以看出，改进后的 FE-RetinaNet 目标检测算法对输入图像中的中小目标能够更加精准的检测和分类。在复杂场景中，与原 RetinaNet 相比，FE-RetinaNet 能够检测到更多的小目标。

3.5　风电场运行检测效果

将 FE-RetinaNet 目标检测算法应用到风电场中，运行效果如图 8 所示。

(a) 原图　(b) 标准RetinaNet　(c) 改进的方法

图 7　MS COCO 数据集主观效果对比

(a) 未佩戴安全帽1　(b) 未佩戴安全帽2　(c) 表计识别1

(d) 表计识别2　(e) 未佩戴安全帽3　(f) 未佩戴安全帽4

图 8　新能源风电场运行效果

从检测结果可以看出,FE-RetinaNet 目标检测算法现场应用效果良好,满足新能源智慧电站对提升设备状态感知能力和自动化巡视能力的实际需要,实现了跨专业协同创新,探索解决传统问题的新思路,运用人工智能的创新驱动力,进一步推动新能源风电场运维自动化技术体系升级。

4 结论

针对 RetinaNet 对于小目标检测效果差的问题,本文提出了改进的 FE-RetinaNet 目标检测算法。首先针对小目标在高层特征图中特征丢失的问题,本文引入了一种主干网改进方案。此外,本文还构造了一个并行的特征增强模块来产生多尺度的上下文特征。然后将特征增强分支与改进后的主干网结合,形成了双向特征金字塔结构,以加强浅层特征的传递效率。将算法在 MS COCO 数据集上进行测试,实验结果表明:改进后的方法检测效果得到了很大的改善,在 COCO 数据集上的 mAP 提高了 1.8%;融入的辅助多尺度特征增强模块有效提高了小目标的检测效果,在 COCO 数据集上 AP_s 提高了 3.3%。并满足在风电场的运行要求,推动新能源风电场运维自动化水平。

参考文献

[1] GIRSHICK R, DONAHUE J, DARRELL T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2013: 1311-1334.

[2] GIRSHICK R. Fast R-CNN[C]//Proceedings of the IEEE International Conference on Computer Vision. Piscataway: IEEE, 2015: 1440-1448.

[3] 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, IEEE, 2017, 39(6): 1137-1149.

[4] HE K, GKIOXARI G, DOLLÁR P, et al. Mask R-CNN[C]//Proceedings of the IEEE International Conference on Computer Vision. Piscataway, NJ: IEEE, 2017: 2961-2969.

[5] DAI J, LI Y, HE K, et al. R-fcn: Object detection via region-based fully convolutional networks[C]//International Conference on Neural Information Processing Systems. New York: ACM, 2016: 379-387.

[6] LIU W, ANGUELOV D, ERHAN D, et al. Ssd: Single shot multibox detector [C]//European Conference on Computer Vision. Springer International Publishing. Berlin: Springer, 2016: 21-37.

[7] REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: unified, real-time object detection [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2016: 779-788.

[8] REDMON J, FARHADI A. YOLO9000: better, faster, stronger[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2017: 7263-7271.

[9] REDMON J, FARHADI A. Yolov3: an incremental improvement [C]//IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 3375-3386.

[10] BOCHKOVSKIY A, WANG C Y, LIAO H Y M. Yolov4: optimal speed and accuracy of objectdetection[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2020: 629-640.

[11] LIN T Y, GOYAL P, GIRSHICK R, et al. Focal loss for dense object detection[C]//Proceedings of the IEEE International Conference on Computer Vision. Piscataway: IEEE, 2017: 2980-2988.

[12] 宋晓茹,杨佳,高嵩,等. 基于注意力机制与多尺度特征融合的行人重识别方法[J]. 科学技术与工程,2022,22(4):1526-1533.

[13] LIN T Y, DOLLÁR P, GIRSHICK R, et al. Feature pyramid networks for object detection[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2017: 2117-2125.

[14] IMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[J]. Computer Science, 2014, 12(89): 936-944.

[15] HE K, ZHANG X, REN S, et al. Identity mappings in deep residual networks [C]//European Conference on Computer Vision. Berlin: Springer, 2016: 630-645.

[16] 刘晋川,黎向锋,叶磊,等. 基于改进 RetinaNet 的行人检测算法[J]. 科学技术与工程,2022,22(10):4019-4025.

[17] 林轶,陈琳,王国鹏,等. 改进的 YOLOv3 交通标志识别算法[J]. 科学技术与工程,2022,22(27):12030-12037.

[18] 梁礼明,尹江,彭仁杰,等. 基于多尺度注意力的皮肤镜图像自动分割算法[J]. 科学技术与工程,2021,21(34):14644-14650.

[19] 赵辉,姜立锋,王红君,等. 基于机器视觉的指针式仪表检测[J]. 科学技术与工程,2021,21(34):14665-14672.

[20] 郑伟,杨晓辉,吕中宾,等. 基于改进 YOLOv4 输电线关键部件实时检测方法[J]. 科学技术与工程,2021,21(24):10393-10400.

[21] 陈亚晨,韩伟,白雪剑,等. 基于改进 YOLO-v3 的眼机交互模型及实现[J]. 科学技术与工程,2021,21(3):1084-1090.

[22] WU A, ZHENG W, YU H, et al. RGB-Infrared Cross-Modality Person Re-identification [C]//IEEE International Conference on Computer Vision (ICCV). Piscataway, NJ: IEEE, 2017: 5390-5399.

Small Target Detection with Parallel Multi-scale Feature Enhancement

HOU Xiaohui[1], LI Li[2], SUN Hongkai[1], MA Liang[1]

(1. Datang Group Co., Ltd., Chifeng 024000, Inner Mongolia, China;
2. College of New Energy, North China Electric Power University, Beijing 102206, China)

Abstract: Because small targets have fewer pixels and carry fewer features, most target detection algorithms can not effectively use the edge information and semantic information of small targets in the feature map, resulting in low precision of small target detection, and the phenomena of missed detection and false detection occur from time to time. In order to solve the defect of insufficient information features of small targets in RetinaNet model, a parallel assisted multi-scale feature enhancement module MFEM (muti scale feature enhancement model) in RetinaNet model is introduced. By using hole convolution with different expansion rates, it avoids information loss caused by multiple down sampling, and is conducive to assisting in shallow extraction of multi-scale context information. In addition, a backbone improvement scheme specially designed for target detection task is adopted, which can effectively save the small target information of high-level feature map. The traditional top-down pyramid structure focuses on transferring high-level semantics from top to bottom, and the one-way information flow is not conducive to the detection of small targets. The auxiliary MFEM branch with RetinaNet is combined to construct a model containing a bidirectional feature pyramid structure, which can effectively integrate the high-level strong semantic information and the low-level high-resolution information. In order to prove the effectiveness of the proposed algorithm FE-RetinaNet (Feature Enhancement RetinaNet), experiments are carried out on MS COCO public data set. Compared with the original RetinaNet, the detection accuracy (mAP) of the improved RetinaNet on MS COCO dataset has been improved by 1.8%, and the COCO AP is 36.2%. FE-RetinaNet has a good detection effect on small targets, and AP_s has increased by 3.2%.

Keywords: target detection; small target; feature enhancement