文章编号：1005-5630(2022)05-0014-06 **DOI**: 10.3969/j.issn.1005-5630.2022.005.002

基于改进特征金字塔的小目标增强检测算法

瑚 琦[1,2]，卞亚林[1,2]，王 兵[1,2]
（1. 上海理工大学 光电信息与计算机工程学院，上海 200093；
2. 上海理工大学 上海市现代光学系统重点实验室，上海 200093）

摘要：小尺寸的物体由于其在图像中分辨率相对较低的原因，在检测任务中容易被丢失和误判。针对目前目标检测算法对小尺寸目标检测精确度远低于其他尺寸目标检测精度的问题加以改进，将小尺寸目标特征增强融入特征金字塔结构。利用多尺度特征融合的特征增强能力丰富小尺寸目标特征层的特征信息，从而使小尺寸目标检测精准度得到提升。将改进特征金字塔结构应用于 YOLOv3 网络，实验对比研究表明，小尺寸目标检测精准度可以达到 0.179，较原网络提升了 22.6%。

关键词：特征金字塔；小目标检测；特征增强；特征融合
中图分类号：TP 391 **文献标志码**：A

Small object enhancement detection algorithm based on improved feature pyramid

HU Qi[1,2]，BIAN Yalin[1,2]，WANG Bing[1,2]
（1. School of Opitcal-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China;
2. Shanghai Key Laboratory of Mordern Optical Systems, University of Shanghai for Science and Technology, Shanghai 200093, China）

Abstract: Small objects are easy to be lost and misjudged in the detection task because of their relatively low resolution in the image. Aiming at the problem that the detection accuracy of small-scale targets in the current target detection algorithm is much lower than that of other sizes, the feature enhancement of small-scale targets is integrated into the feature pyramid structure to avoid the lack of small-scale feature information. The feature enhancement ability of multi-scale feature fusion is used to enrich the feature information of small-scale target feature layer, so as to improve the accuracy of small-scale target detection. The improved feature pyramid structure is applied to YOLOv3 network. The experimental comparative study shows that the detection accuracy of small-

收稿日期：2022-03-06
基金项目：国家自然科学基金（61975125）
作者简介：瑚 琦 (1977—)，男，副教授，研究方向为光电检测与仪器设计、嵌入式人工智能。E-mail：harehuqi@163.com
通信作者：卞亚林 (1996—)，男，硕士研究生，研究方向为目标检测。E-mail：192380303@st.usst.edu.cn

scale targets can reach 0.179, which is 22.6% higher than the original network.

Keywords: feature pyramid network; small object detection; feature enhancement; feature fusion

引　言

目标检测是数字图像处理的基础课题之一，同时也是计算机视觉领域的重要分支[1]。目标检测的任务是确定图像中目标的类别并给出位置信息。目标检测对特征信息的提取及语义信息的应用都有很高要求，高质量的检测结果对更为复杂的计算机视觉任务的完成有着至关重要的作用。

小目标检测是目标检测的一大难点，在医疗检测、自动驾驶和工业检测中都有着广阔的应用需求和前景。因此，提升小目标的检测效果有很强的现实需求。近年来，随着深度学习和卷积神经网络在计算机视觉领域取得重大突破，小目标检测也逐渐成为研究热点。

根据国际光学工程学会的定义，图像中小于原图尺寸 0.12% 的物体即为小目标。在通用目标检测数据集（microsoft common objects in context，MS COCO）[2] 中，小于 32 像素×32 像素的物体会被定义为小目标。基于深度学习实现小目标检测存在的难点主要有 3 点：(1) 如何将小目标与其他目标的检测区分开来；(2) 通用目标检测数据集中小目标数目相对较少；(3) 小目标包含的特征信息较少。因此，基于深度学习实现的小目标检测算法的研究现状，可以总结为 3 个方向。

多尺度检测，图像金字塔尺度归一化（scale normalization for image pyramids，SNIP）[3] 借鉴图像金字塔，对图像金字塔中不同尺度图像分别训练检测器，再合并检测结果以实现多尺度目标检测。Liu 等[4] 提出的一次多框检测器（single shot multibox detector，SSD）直接利用主干网络的不同尺度特征层负责不同尺度物体的检测。Redmon 等[5] 在只看一次目标检测算法（you only look once version 3，YOLOv3）引入了特征金字塔（feature pyramid networks，FPN）[6] 结构丰富特征信息，并通过设置不同大小的锚框实现多尺度物体的检测效果。

数据增强，Mixup 方法[7] 通过随机选取训练集中两张图像进行加权求和的方式进行数据集扩展。Cutmix[8] 以及 Mosaic[9] 两种方法分别混合两个图像和四个图像来增强数据集的表达能力。Copy-pasting 增强方法[10]则通过复制小目标的方式丰富小目标的数目和分布情况。

特征增强，通道增强网络（path aggregation network，PANet）[11] 在特征金字塔后通过一个自下而上的路径增强缩短低层级特征和特征图的连接，增强了特征金字塔。卷积注意模块（convolutional block attention module，CBAM）[12] 即尺度变换模块，通过池化层获得小尺度特征图，通过尺度转换层减少特征图通道数以获得大尺度的特征图。平衡特征金字塔网络（libra feature pyramid networks，Libra-FPN）[13] 将不同特征层做尺度变换并进行特征融合以增强特征金字塔的尺度表达平衡。

在特征金字塔结构中，通过水平融合低分辨率特征图和高分辨率特征图的方式虽然可以缓解多尺度目标检测不同特征层间信息扩散的问题，但同时也会导致语义冲突，使小目标的特征信息被稀释。因此，本文以 YOLOv3 为基础网络，提出通过跳跃连接和多尺度特征融合的方式，丰富网络提取的特征信息。通过在 MS COCO 2017 数据集上的实验表明，改进后的网络在小尺度目标检测上的平均精度均值（mean average precision，mAP）较原网络提升了约 3%。

1　改进的特征金字塔结构

1.1　多尺度特征融合模块

FPN 通过卷积神经网络前向过程形成的具有金字塔结构的特征层，加上一个带有横向链接的从上而下结构，既能丰富网络中获取的特征信息和语义信息，也能实现多尺度目标检测任务。

一般认为，卷积神经网络前向过程中，由于卷积核的作用，深层网络中的节点具有较大的感受野，因此相较于浅层网络，包含的语义信息更

多，特征信息更少。因此，为了使多尺度目标检测网络中不同尺度目标获得更多的特征信息和语义信息，特征融合是常用的手段。不同的特征融合方式会产生不同的效果。现阶段通过特征融合实现特征增强的方法大多都是将浅层特征信息和深层特征信息分别融合，以实现多尺度目标检测效果均有提升，但会增加网络计算负担，并且多尺度目标检测精确度之间的不平衡现象没有得到改善。

为实现通过较少计算量提升小尺度目标检测精确度的目的，本文通过如图 1 所示的特征融合模块，使原 F5 特征层中由于主干网络前向过程丢失的特征信息获得直接的增补，从而对小目标的检出提供特征信息上的帮助。特征融合阶段没有需要额外学习的参数，所有特征信息均从原主干网络提取并进行尺度变换获得。

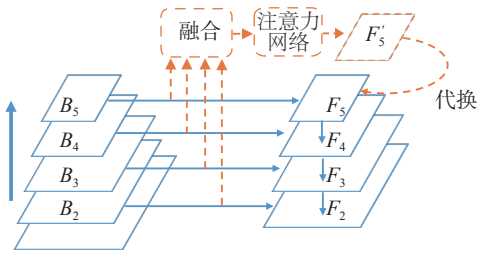


图 1　改进的特征金字塔结构

Fig. 1　Improved feature pyramid network

特征融合过程主要包含 3 个步骤：

（1）通过池化可以实现特征层尺度的统一。利用 1×1 卷积核，可以通过控制卷积核的个数调整特征层的通道数，使得它们保持一致。这个过程可以表示为：

$$B_n' = F_{\text{rechannel}}\left(F_{\text{maxpool}}\left(B_n\right)\right) \tag{1}$$

式中：B_n 为从主干网络中提取的第 n 层特征层；$F_{\text{rechannel}}$ 代表通道数调整函数；F_{maxpool} 代表最大池化函数；B_n' 为调整后用于特征融合的特征层。

（2）将尺度和通道数均一致的特征层直接进行特征融合。这个过程可以表示为：

$$P_5' = \frac{1}{N-1}\sum_{n=2}^{N} B_n' \tag{2}$$

（3）将融合后的特征，运用注意力模块（SENet[14]）调整注意力，以增强特征图表达能力。SENet 主要包括 Squeeze 和 Excitation 两个操作：

$$z = F_{\text{sq}}(P_5) = \frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W} P_5(i,j) \tag{3}$$

$$s = F_{\text{ex}}(z,W) = \sigma(g(z,W)) = \sigma(\boldsymbol{W}_2\delta(\boldsymbol{W}_1 z)) \tag{4}$$

Squeeze 操作将某一通道上的全部空间特征编码为一个全局特征，然后使用 Sigmoid 激活函数和 ReLU 激活函数对通道特征编码信息进行重新校正。式 (4) 中：σ 表示 ReLU 激活函数；δ 表示 Sigmoid 激活函数；$\boldsymbol{W}_1$ 和 $\boldsymbol{W}_2$ 表示两个用来限制模型复杂度的权重参数矩阵。

将式 (4) 提取到的激活值乘上式 (2) 中融合后的特征信息，得到最终改进后的特征层：

$$F_5' = s\cdot P_5' \tag{5}$$

1.2　特征融合模块在 YOLOv3 中的应用

YOLOv3 是由 YOLO 系列[15-16] 发展而来，是单阶段目标检测网络中的经典模型。YOLOv3 主干网络采用 Darknet53，在原 Darknet19 上结合残差网络结构思想，使用连续的 3×3 和 1×1 卷积层构成一个共包含 53 个卷积层的主干网络。YOLOv3 借鉴了特征金字塔的思想，用具有更大感受野的小尺寸特征图检测大尺寸的物体，而具有相对较小感受野的大尺寸特征图检测小尺寸物体。以预处理将输入图片的尺寸调整成 416 像素×416 像素为例，则从主干网络提取 3 种尺度的特征图，分别为 52×52、26×26、13×13 像素值。然后，再引入 9 种不同大小的先验框，分别给每种尺度特征图分配 3 种，从而实现了多尺度的目标检测。

YOLOv3 引入残差网络结构的目的是为了提取更深层次的语义信息，但更深的网络结构也会造成不同尺度特征图之间特征信息提取的差异，从而造成不同尺度目标检测精确度之间的差异。以网络输入尺寸为 416 像素×416 像素为例，依据 MS COCO数据集的标准，YOLOv3 检测网络对大中小尺寸目标的检测精度分别为 44.8%、33.6% 和 14.6%，可以看出小目标的检测精确度远低于其他尺寸。因此，将针对小目标增强的特征金字塔结构应用于 YOLOv3，如图 2 所示。

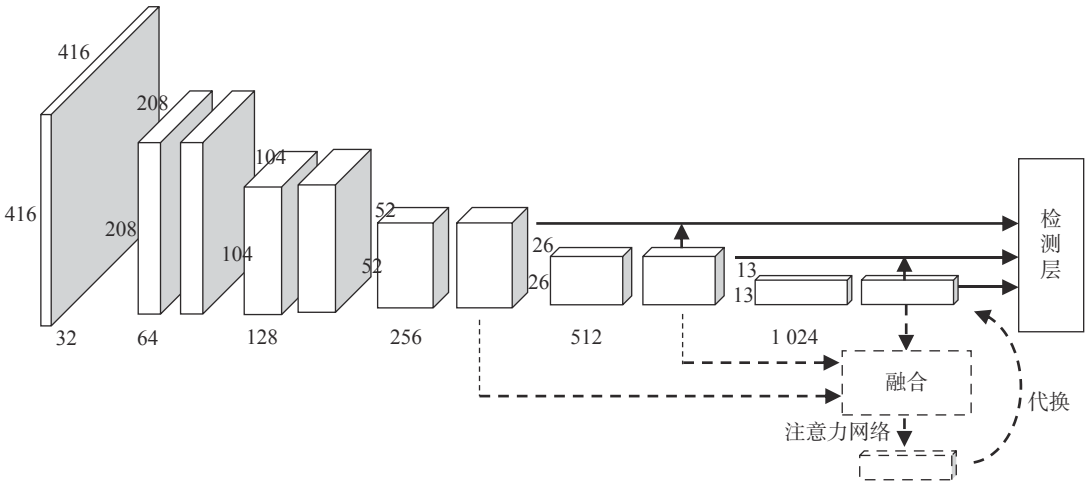


图 2　特征增强 YOLOv3 示意图

Fig. 2　Schematic diagram of feature enhanced YOLOv3

2　实验结果及分析

2.1　实验设置

本文实验设备为 2.9 GHz，i5-9400F CPU，GeForce RTX 3060 GPU，16 G RAM 的计算机上运算。操作系统为 Ubuntu18.04，软件运行环境 PyTorch1.6.0，CUDA10.1。训练设置中，在数据集上的实验进行 275 轮训练，设置 batch_size 为 8，初始学习率为 0.002。

2.2　数据集及评价指标

MS COCO 数据集是目前通用目标检测研究领域广泛使用的一个大规模数据集，包含 80 个检测类在内的 33 万张图片。本文选取的 COCO 数据集 2017 版本，共包含训练集图片 118 287 张，验证集图片 5 000 张。

在 MS COCO 检测任务中，主要的评估指标是平均精度均值 *mAP*。*mAP* 是对每个类别的目标的平均精度 *AP* 取均值，而 *AP* 是由准确率−召回率（precision-recall）曲线围成的面积决定的。因此，需要分别计算准确率和召回率。

假设分类的目标有两类，正类（positive）和负类（negative），通过网络输出的置信度值（confidence）判定。通过设定 IoU 阈值判定预测结果是否正确，通常设定的阈值为 0.5，不同数据集的评价标准会有浮动。据此，检测结果可以被分为 4 类：（1）被预测为正类的正类样本（true positive，TP）；（2）被预测为正类的负类样本（false positive，FP）；（3）被预测为负类的正类样本（false negative，FN）；（4）被预测为负类的负类样本（true negative，TN）。则准确率和召回率的计算公式分别为：

$$Precision = \frac{\text{被正确预测的正类样本数}}{\text{所有预测为正类的样本数}} = \frac{TP}{TP+FP} \tag{6}$$

$$Recall = \frac{\text{被正确预测的正类样本数}}{\text{所有正类样本数}} = \frac{TP}{TP+FN} \tag{7}$$

通过式 (6) 和式 (7) 得到图 3 所示准确率−召回率曲线，计算曲线围成的面积得到某类目标的 *AP*，再根据式 (8) 可计算得到最终的 *mAP*。

$$mAP = \frac{\sum AP}{N_{\text{class}}} \tag{8}$$

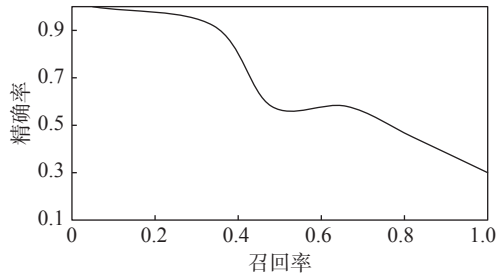


图 3　精确率−召回率曲线示意图

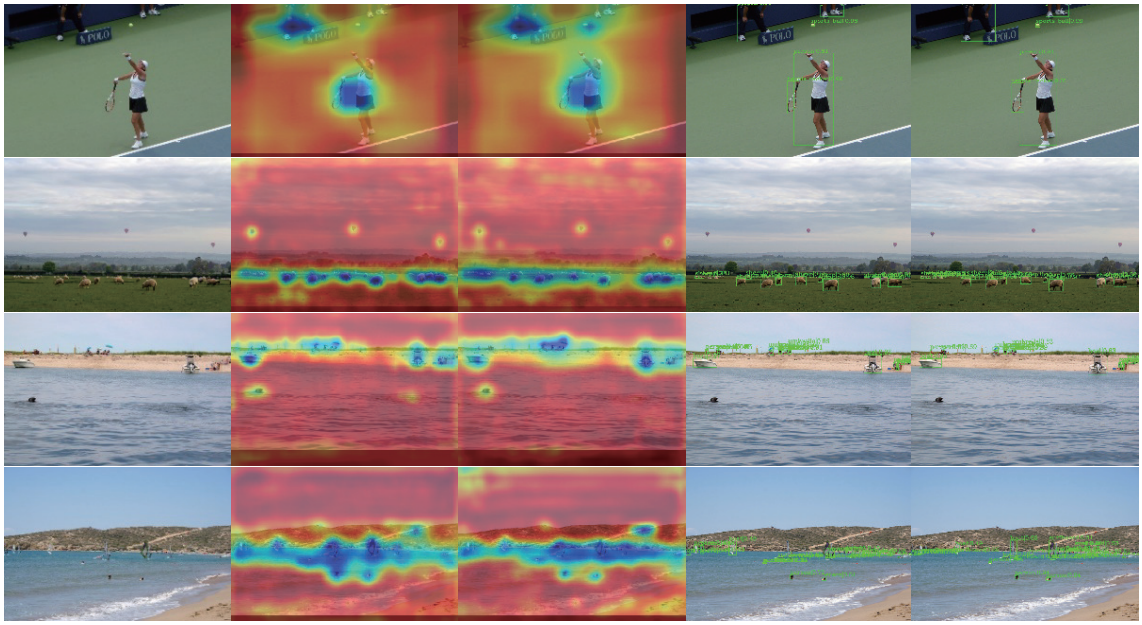
Fig. 3　Figure of precision-recall curve

针对 COCO 数据集中不同尺寸目标较多，分别对大、中、小 3 类尺度的检测精度进行评价，以衡量模型对不同尺度物体的检测效果。其

中，AP_S 指原图像中尺寸小于 32 像素×32 像素的目标的 *mAP*；AP_M 指原图像中尺寸在 32 像素×32 像素与 96 像素×96 像素之间的目标的 *mAP*；AP_L 指原图像中尺寸大于 96 像素×96 像素的目标的 *mAP*。

2.3　实验结果分析

为验证本文的改进特征金字塔结构在小目标检测特征提取和精确度上的变化，将本文网络与改进前的 YOLOv3 网络及其他目标检测网络进行比较。在设定网络的输入尺寸为 416 像素×416 像素的情况下，本文网络与改进前的 YOLOv3 网络提取小尺寸目标特征情况与检测结果如图 4 所示。图 4 中，(a) 列为包含小目标的原图，(b) 列和 (c) 列分别为模型改进前后输出的特征图，(d) 列和 (e) 列分别为模型改进前后输出的检测结果。从特征图的对比中可以看出，本文改进后的网络能够提取更丰富的小尺寸目标信息，从检测结果对比中，小尺寸目标的检测结果在本文模型上取得更高的精确度。如图 4 中第一行中，网球属于需要被检测的小目标，可从图中看出，改进后的检测网络能够获得更多的网球特征信息，也能够将网球的检测精确度由 0.96 提升至 0.99。



(a) 检测原图　(b) 原网络特征图　(c) 本文特征图　(d) 原网络检测结果　(e) 本文检测结果

图 4　本文网络与原 YOLOv3 实验结果对比图

Fig. 4　Schematic diagram of feature enhanced YOLOv3

由于本文侧重提升小目标检测的精确度，因此选取交并比（IoU）为 0.5 的阈值以及小尺寸目标检测精度 AP_S 作为评价指标，如表 1 所示。

表 1　MS COCO2017 数据集精确度测试结果

Tab. 1　*mAP* on MS COCO2017 dataset

方法	主干网络	*mAP*(%),IoU:0.5	*mAP*(%),S
YOLOv2	Darknet-19	43.9	5.3
YOLOv3	Darknet-53	55.3	14.6
Faster R-CNN+++	ResNet-101	55.7	15.6
Our approach	Darknet-53	56.7	17.9

从表 1 中可以看出，将改进特征金字塔结构应用于 YOLOv3检测网络，可以将网络整体检测精确度提升1.4%，小目标检测精确度提升 3.3%。小目标检测相比 YOLOv3 的精度提高了 22.6%。与使用更复杂主干网络的二阶段目标检测网络 Faster R-CNN 相比，本文网络结构的检测精确度也有所提高。

综合以上情况，本文提出的基于改进特征金字塔的小目标增强检测网络，可以增强检测网络对小尺寸目标的特征提取，从而提升其检测精确

度，并有助于目标检测网络的整体提升。

3　结　论

针对小目标包含特征信息少，检出精确度差的特点，本文提出了一种基于改进特征金字塔的 YOLOv3 目标检测网络，通过一个多尺度特征融合模块取代原特征金字塔中局部特征层，使网络获得更丰富的小目标特征信息。与原算法模型相比，本文提出的小目标增强目标检测算法在增加较少计算量的基础上，增加了小目标物体特征检出效果，小目标检测精确度提升了 3.3%。但本文算法还存在一定改善空间，如模型未考虑小目标被局部遮挡的情况。由于小目标本身包含像素信息较少，遮挡问题对小目标的检出效果会产生很大影响，因此后续工作可以对算法模型作进一步优化，提升网络在不同应用场景的检测性能。

参考文献：

[1] 南晓虎, 丁雷. 深度学习的典型目标检测算法综述[J]. 计算机应用研究, 2020, 37(S2): 15 – 21.

[2] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft COCO: common objects in context[C]//13th European Conference on Computer Vision. Zurich: Springer, 2014: 740-755.

[3] SINGH B, DAVIS L S. An analysis of scale invariance in object detection - SNIP[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 3578-3587.

[4] LIU W, ANGUELOV D, ERHAN D, et al. SSD: single shot MultiBox detector[C]//14th European Conference on Computer Vision. Amsterdam: Springer, 2016: 21-37.

[5] REDMON J, FARHADI A. YOLOv3: An incremental improvement[DB/OL]. (2018-04-08). https://arxiv.org/abs/1804.02767.

[6] 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. Honolulu: IEEE, 2017: 936-944.

[7] ZHANG H Y, CISSÉ M, DAUPHIN Y N, et al. mixup: beyond empirical risk minimization[C]//6th International Conference on Learning Representations. Vancouver: OpenReview. net, 2018.

[8] YUN S, HAN D, CHUN S, et al. CutMix: regularization strategy to train strong classifiers with localizable features[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019: 6022-6031.

[9] BOCHKOVSKIY A, WANG C Y, LIAO H Y M. YOLOv4: Optimal speed and accuracy of object detection[DB/OL]. (2020-04-23). https://arxiv.org/abs/2004.10934.

[10] KISANTAL M, WOJNA Z, MURAWSKI J, et al. Augmentation for small object detection[DB/OL]. (2019-02-19). https://arxiv.org/abs/1902.07296.

[11] LIU S, QI L, QIN H F, et al. Path aggregation network for instance segmentation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 8759-8768.

[12] Woo S, Park J, Lee J Y, et al. Cbam: Convolutional block attention module[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 3-19.

[13] PANG J M, CHEN K, SHI J P, et al. Libra R-CNN: towards balanced learning for object detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 821-830.

[14] HU J, SHEN L, ALBANIE S, et al. Squeeze-and-excitation networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 42(8): 2011 – 2023.

[15] 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. Las Vegas: IEEE, 2016: 779-788.

[16] REDMON J, FARHADI A. YOLO9000: better, faster, stronger[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017: 6517-6525.

（编辑：李晓莉）