• 成电青年学者·信息与通信工程专栏 •

基于空洞卷积金字塔的目标检测算法

候少麒[1]，梁 杰[1]，殷康宁[1]，刘学婷[1]，殷光强[2*]

(1. 电子科技大学信息与通信工程学院 成都 611731；2. 电子科技大学信息与软件工程学院 成都 610054)

【摘要】作为目标检测领域最突出的问题，遮挡和多尺度严重影响了算法的召回率和准确率。针对以上问题，该文从感受野入手，提出了一种基于空洞卷积金字塔网络 (ACFPN) 的目标检测算法。首先，将不同尺寸的空洞卷积层引入特征金字塔网络 (FPN) 中，构建混合感受野模块 (HRFM)，旨在控制参数量的条件下，通过增大感受野获取更多全局特征信息，解决目标的遮挡问题；其次，改进 FPN 的结构，设计低层嵌入特征金字塔模块 (LEFPM)，将浅层特征细节信息和高层特征语义信息相融合，提高特征图的丰富度和表征能力，增强模型的尺度适应性；特别地，针对漏检问题，引入 FCOS 算法中的无锚框 (AF) 机制，减少了候选框的冗余，进一步提高了定位精度。最后在公开数据集上进行测试，该算法在检测精度上大幅提升。

关 键 词 空洞卷积； 特征融合； 特征金字塔； 目标检测； 感受野

中图分类号 TP391.4 **文献标志码** A doi:10.12178/1001-0548.2021032

Object Detection Algorithm Based on Atrous Convolutional Pyramid

HOU Shaoqi[1], LIANG Jie[1], YIN Kangning[1], LIU Xueting[1], and YIN Guangqiang[2*]

(1. School of Information and Communication Engineering, University of Electronic Science and Technology of China Chengdu 611731;
2. School of Information and Software Engineering, University of Electronic Science and Technology of China Chengdu 610054)

Abstract As the most prominent problem in the field of object detection, occlusion and multi-scale seriously affect the recall and precision of the algorithm. To resolve the problems mentioned above, this paper starts from the receptive field, proposing an object detector based on the atrous convolution embedded feature pyramid network (ACFPN). Firstly, the atrous convolutional layers of different sizes are introduced into the feature pyramid to construct a hybrid receptive field module (HRFM), which aims to obtain more global feature information by increasing the receptive field with the number of parameters staging roughly the same, thereby solving the problem of occlusion; secondly, by improving the structure of the feature pyramid, we design a lower embedding feature pyramid module (LEFPM) to enhance the model's scale adaptability, which combines shallow feature's detail information and high-level feature's semantic information to improve the richness and representation ability of feature maps; in particular, targing at the problem of missed detection, the Anchor Free mechanism of the fully convolutional one-stage (FCOS) algorithm is introduced to reduce the redundancy of candidate frames and further improve the positioning accuracy. The algorithm is tested on the public VOC dataset , and has shown a great improvement on detection accuracy.

Key words atrous convolution; feature fusion; feature pyramid; object detection; receptive field

目标检测是现实生活中最广泛的应用之一，其任务在于关注图片中的特定目标。一般来说，通用性目标检测包含两个子任务：一是判定特定目标的类别概率，二是给出该目标的具体位置。目标检测在实际应用中有着非常重要的作用，可以运用于人脸识别、行人重识别、工业检测、车牌号识别、医学影像等具体场景，涉及安防领域、工业领域、军事领域、交通领域、医疗领域和生活领域等。随着机器学习的蓬勃发展，普通场景下目标检测的精度已经很高，但针对复杂环境下目标数量众多、目标尺度多变、目标遮挡严重等问题，仍是国内外科研人员的研究重点[1]。

收稿日期：2021－01－31；修回日期：2021－06－30
基金项目：国家重点研发计划 (2018YFC0807501)
作者简介：候少麒 (1992－)，男，博士生，主要从事计算机视觉方面的研究.
*通信作者：殷光强，E-mail：yingq@uestc.edu.cn

传统的基于手工特征构建的目标检测算法过程复杂、计算量大，但为目标检测的发展奠定了理论基础。作为传统领域最经典的算法，文献 [2] 的目标检测器通过多尺度滑动窗口来生成可能存在的具有不同宽高比的目标区域，再利用模板进行目标匹配。另外一个与之相似的传统方法是利用梯度直方图 (histogram of oriented gradient, HOG)[3] 特征和支持向量机 (support vector machine, SVM)[4] 来进行目标分类。

随着计算机视觉技术的长足发展，基于深度学习的目标检测开始成为研究热门。在 2012 年 ImageNet 竞赛上取得冠军的 AlexNet[5]，是首个在大规模图像识别问题取得突破性进展的深度神经网络，并由此开启了深度神经网络在计算机视觉领域的广泛应用。基于深度神经网络的目标检测算法按照处理分类和回归的方法差异，又可划分为单阶段 (one stage) 和两阶段 (two stage) 两大派系。

两阶段算法中，以 RCNN[6] 为代表的目标检测算法，其核心是采用区域提议方法，对输入图像进行选择性搜索并生成区域建议框，然后对每一个区域建议框使用卷积神经网络 (convolutional neural networks, CNN) 提取特征，再使用分类器进行分类。该类方法最大的短板是冗余框的重复计算，导致最快的算法[7] 在 GPU 上也只有 7 帧/s 的推理速度。另一类单阶段目标检测算法是以 YOLO[8-10] 和 SSD[11] 为代表的基于直接回归的算法。这类算法将单个神经网络应用于整幅图像，并在最终的特征图上划分网格区域，同时预测每个区域的边界框和目标概率，在牺牲一定精度的同时大大减少了重复计算。

经过一系列的变种，这两类方法的共同点逐渐演变为在检测过程中都需要预先生成大量锚框 (anchor)，这些算法统称为基于锚框 (anchor based) 的目标检测算法。锚框是在训练之前，在训练集上利用聚类算法得出的一组矩形框，代表数据集中目标主要分布的长宽尺寸。在推理时先在特征图上由这些锚框提取 n 个候选矩形框，再对这些矩形框做进一步的分类和回归。相对 Two Stage 算法来说，对候选框的处理依然经过前背景粗分类和多类别细分类两步。

单阶段目标检测算法由于缺少了两阶段算法的精细处理，在面对目标多尺度、遮挡等问题时表现不佳。另外，Anchor Based 算法虽然在一定程度上缓解了选择性搜索带来的候选框计算量爆炸的问题，但每个网格中大量不同尺寸锚框的生成仍然造成了计算冗余，最重要的是锚框的生成依赖于大量的超参设置，手动调参会严重影响目标的定位精度和分类效果。

针对以上问题，本文提出了一种基于空洞卷积金字塔的目标检测算法 (atrous convolution embedded feature pyramid network, ACFPN)，能够有效地解决因尺度和遮挡引起的漏检、错检问题，主要创新点如下：

1) 设计多尺寸的空洞卷积构成的混合感受野模块 (hybrid receptive field module, HRFM)，结合特征金字塔多尺度输出特性，在控制模型参数量的条件下，增大感受野获取更多全局特征细节信息，以解决目标的遮挡问题；

2) 改进特征金字塔网络的结构，提出了低层嵌入特征金字塔模块 (lower embedding feature pyramid module, LEFPM)，解决目标检测在处理多尺度变化上不足，融合浅层特征信息和高层特征信息，并在融合后的输出增加归一化处理和激活函数，优化模型训练；

3) 引入 Anchor Free 机制，结合上述两点设计，减少冗余候选框带来的无效计算，提高了定位精度，有效解决漏检等问题。

1 相关工作

1.1 特征金字塔

很多传统目标检测方法都会使用图像金字塔来解决目标的多尺度问题。图像金字塔首先将不同尺寸的图片分别输入网络中得到对应尺寸的特征图，然后对这些不同尺寸的特征图进行预测。这种方法虽然可以在一定程度上应对尺寸变化，但是带来了成倍的计算量。进入深度学习时代后，目标检测器在精度方面取得了显着提高，文献 [12] 提出了 SPPNet，该算法使用空间金字塔池化策略，对输入任意尺寸图像都能够产生固定大小的特征图。早期基于深度模型的检测器只在网络的顶层进行检测，特征单一且适应性差。文献 [13] 基于 Faster RCNN[7] 提出了特征金字塔网络 (feature pyramid network, FPN)，FPN 具有横向连接的自顶向下体系结构。本文改进特征金字塔网络结构，提出了 LEFPM 模块。两者结构对比如图 1 所示。用于在所有级别特征中构建高级语义信息，由于在检测多尺度目标时效果显著，FPN 已经成为众多深度检测器的标准配置。

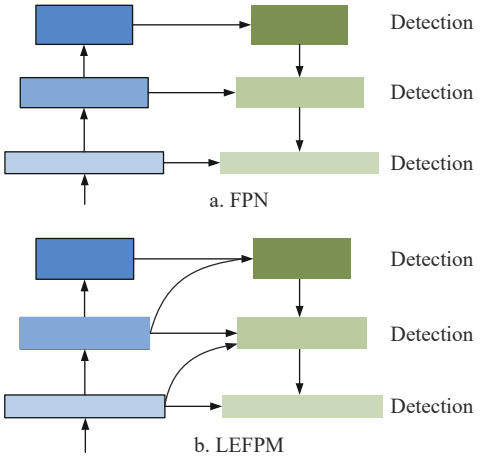


图 1　两种特征金字塔的结构对比

1.2　空洞卷积

在空洞卷积出现之前，研究人员普遍通过降采样方式来间接增大感受野，降采样方式会丢失大量有用信息，还易造成特征图分辨率的急剧下降。2016 年，文献 [14] 在图像分割领域提出了空洞卷积模型，有效解决了这一难题。

空洞卷积是通过在标准卷积中进行零填充的方式，扩大卷积核的尺寸，使其能够更好地捕获特征图的上下文信息。空洞卷积的尺寸通过调整空洞率 (atrous rate, ar) 来实现，空洞率即为在卷积核相邻参数中填充 (ar−1) 个 0。同样的，标准卷积可以看做空洞卷积在 ar 为 1 时的特殊形式。

空洞卷积的输出定义为：

$$y(i,j)=\sum_{h=1}^{H}\sum_{w=1}^{W}x(i+\mathrm{ar}\times h, j+\mathrm{ar}\times w)\times w(h,w) \quad (1)$$

式中，H、W 分别表示输入图像 (或前一层特征图) 的长和宽；$x(i,j)$表示该输入图像上 (i, j) 位置的像素值 (特征值)；ar 表示空洞率；$y(i,j)$表示该输入图像经过空洞卷积后的输出。

空洞卷积可以在不损失特征图分辨率的情况下，有效聚合图像全局特征信息，从而增加其感受野，解决目标的遮挡问题。同时因为其填充值为 0，所以不会增加额外的计算开销。

1.3　Anchor Free 机制

由于密集的锚框可有效提高待测目标的召回率，加之操作简单，现阶段基于 Anchor Based 的目标检测算法依然占据着深度目标检测算法的主流，包括最经典的 Fast R-CNN[15]、SSD[11]、YOLOv2[9]、YOLOv3[10] 等目标检测算法。

然而在基于 Anchor Based 的检测机制中，相关超参的设置严重依赖较强的先验知识。同时，根据预设产生的冗余框非常多，使得正负样本严重不平衡。因此，Anchor Free 方法被越来越多的研究者探索。YOLOv1[8] 在目标中心附近的点处预测边界框实现了 Anchor Free，遗憾的是，其后续版本为了追求高召回率，依然采用了 Anchor Based 路线。在 Anchor Free 算法中，基于关键点的方法 (如 CornerNett[16] 和 CenterNet[17]) 本质上都是密集预测的手段，庞大的解空间使得简单的 Anchor Free 方法容易得到过多的误检，而获得高召回率、低精确率的检测结果。

FCOS[18] 方法从像素点入手，一方面通过重新赋予权重来提高检测质量，另一方面通过加入 FPN 在一定程度上缓解了高度重合带来的影响。

2　基于空洞卷积金字塔的目标检测算法

2.1　整体框架

本文的 ACFPN 算法以一阶段全卷积目标检测算法 FCOS 为基准 (Baseline)，并引入了 FCOS 特有的 Achor Free 机制。ACFPN 主要由 4 部分组成：主干网络、LEFPM、HRFM、检测模块，如图 2 所示。其中，LEFPM 和 HRFM 两个模块都作用于主干网络所产生的特征图，并在整个架构中执行不同的功能。

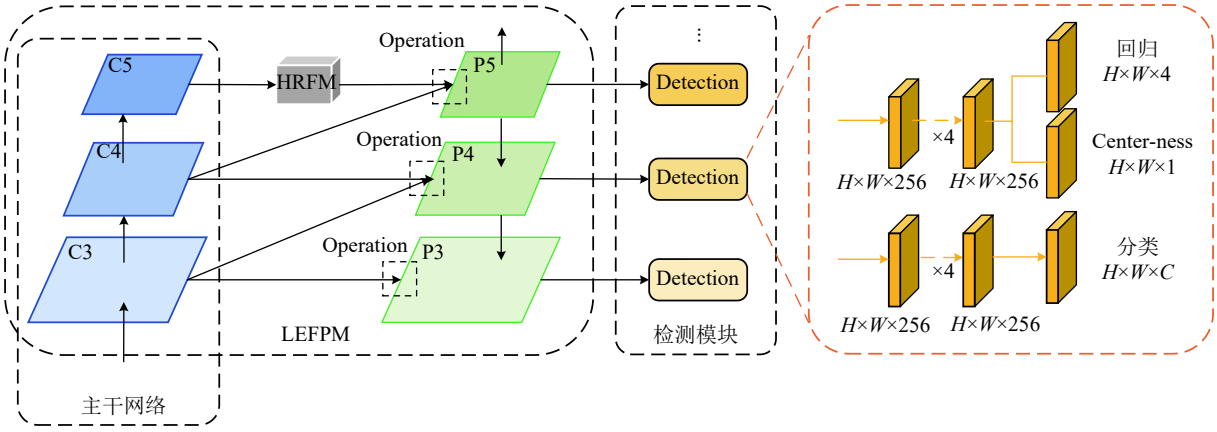
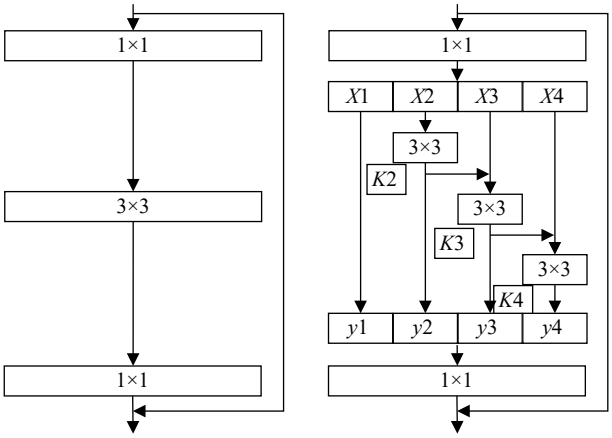


图 2　本文 ACFPN 的整体结构

正如图 2 的结构所示：首先，主干网络对待检测图片进行卷积处理，生成具有不同高、低级特征信息的特征图；其次，顶层特征图 C5 经过 HRFM 进一步处理，并和其他特征图一起送入 LEFPM 中；然后，LEFPM 通过一系列细节操作，将特征图的高、低层信息进行充分融合，并将融合后的特征图输入到最后的检测模块中；最后，检测模块借助不同尺度子网络的组合设计，对不同尺度的待检测目标分别进行分类和定位。

特别地，本文主干网络采用新的多尺度结构 Res2Net50[19] 替换原 FCOS 的 ResNet50。相比于 ResNet50，Res2Net50 在给定冗余块中使用了分层级联特征组取代了通常的单个 3×3 卷积层，如图 3 所示，该特征组在网络宽度、深度和分辨率等方面有更多的优化。



a. ResNet 模块　　b. Res2Net 模块

图 3　ResNet 和 Res2Net 主干网络对比

2.2　混合感受野模块 (HRFM) 设计

众多研究表明，使用单一尺寸的空洞卷积会引起网格效应[20]，即空洞率过大时，卷积会因为插入 0 值而导致过于稀疏，难以捕获关键信息，不利于小目标检测。

为充分利用密集矩阵的高计算性能，文献 [21] 率先提出用密集成分来近似或者代替最优的局部稀疏结构。2017 年是空洞卷积和多尺度密集结构快速发展的一年：为减少信息损失，文献 [22] 提出使用不同尺寸的卷积层特征图融合成具有全局信息的特征表示方法；文献 [23] 模拟人类视觉的感受野，通过在 InceptionNet[21] 中引入空洞卷积，加强网络的特征提取能力；文献 [24] 在人脸检测中也将多尺度密集连接引入上下文信息模块，以解决人脸的遮挡问题。

受以上思想的启发，本文设计了 HRFM 通过将不同空洞率的空洞卷积层并行获取的特征图拼接在一起，提高网络对全局特征的获取能力，弥补单一空洞卷积引起的网格效应。不同于 InceptionNet 和 RFBNet，本文的 HRFM 全部使用空洞卷积层。

经过大量实验，发现空洞卷积对顶层特征图的感受野影响最大，为了充分发挥 HRFM 的性能，特别将 HRFM 嵌入在 C5 和 P5 之间。

由图 4 所示，HRFM 由 4 个分支组成，一个 1×1 的卷积层分支，3 个空洞率分别为 ar=1, 2, 4 的 3×3 卷积层分支。ar=4 的 3×3 空洞卷积层能够获取更多全局性的上下文特征细节，增强推理能力，解决目标遮挡问题；不同空洞率的卷积层使用，提高了模型对不同尺度目标的适应性；特别地，在拼接后的特征图后，采用 1×1 的卷积层进行特征信息融合，并将通道维度降低至指定数量，提高了 HRFM 模块的灵活性。

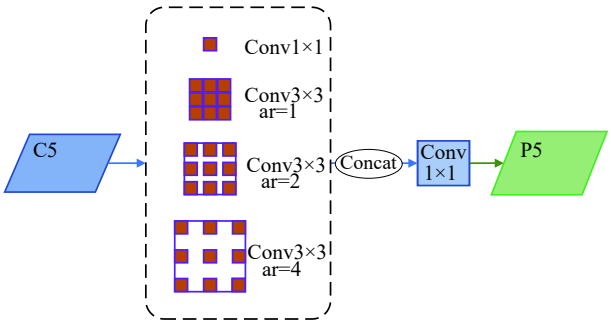


图 4　HRFM 的结构

1×1 的卷积层可以在不改变特征图尺寸的情况下，尽可能地保留图像的细节信息，同时可以控制特征图的通道数，减少后续的计算量；3×3 的卷积层具有较小的参数，既可以对特征信息进行加工，又进一步减少了网络的计算；空洞卷积能够获取更多全局特征细节信息，增强推理能力，对遮挡目标有很好的识别能力，不同空洞率的设置在消除网格效应的同时，也提高了模型对多尺度目标的适应能力。

2.3　低层嵌入式特征金字塔模块 (LEFPM) 设计

研究发现，单阶段目标检测算法无法用单一维度特征图同时有效地表征各个尺度的物体，这类的目标检测算法通常只采用顶层特征做预测。FPN 结构提出具有横向连接的自顶向下架构，虽然将较高层的语义信息引入到当前层特征图，但组合后的复合特征图仍然存在两个问题：

1) FPN 构建用于检测的特征图时，并未考虑来自较低层的特征信息。较高层特征图虽然包含更强的语义信息，但由于被多次下采样和上采样，包

含的位置信息大量缺失。而较低层特征含有更精细的信息，这对于定位和检测小尺度物体很有帮助；

2) FPN 产生的复合特征图既作为高层语义信息向下传递，同时又用于检测，这样使用复合特征图并不合理，因为复合特征图承担了过多的任务。

本文提出的 LEFPM 在 FPN 的基础上，通过低层嵌入的方式，进一步充分融合低层细节信息，以实现对多尺度目标检测效果和定位精度的双提升。

如图 5 所示，其中 C5′是 C5 经过 HRFM 处理后的特征图，复合卷积层由 3×3 卷积层、BN 层和 LeakyReLU 激活层级联而成，目的是加工融合后的特征、优化模型训练，并提高特征的非线性表达能力。

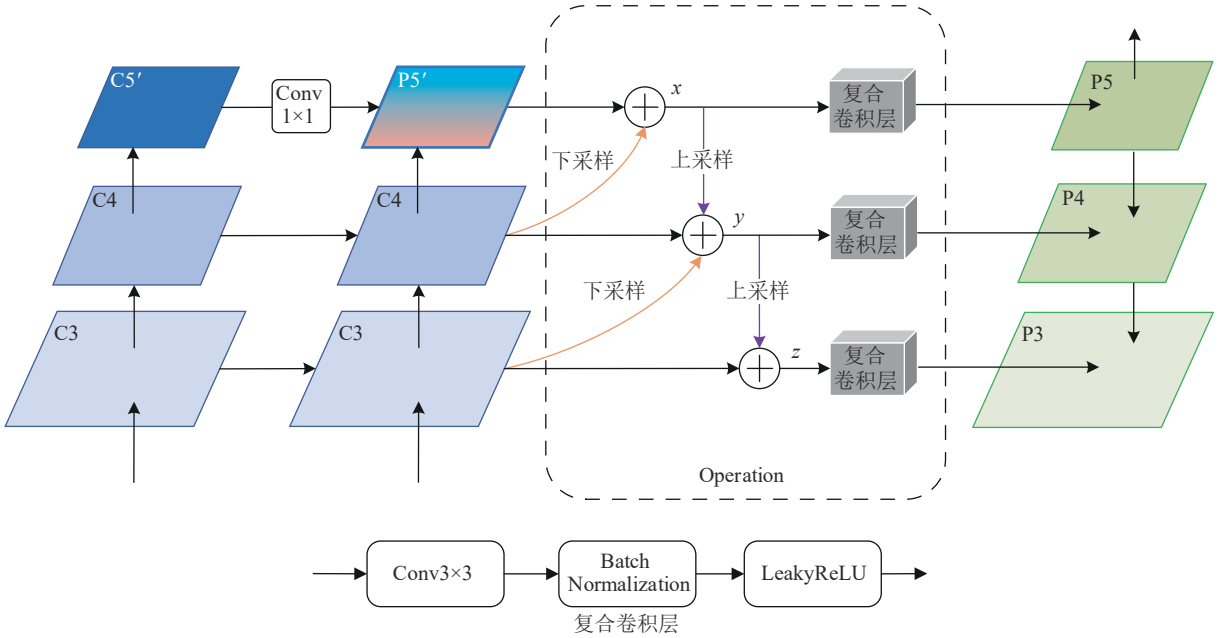


图 5　本文 LEFPM 结构图

LEFPM 的设计思想是，首先将当前层特征图与经过通道压缩和上采样后的高层特征图相融合(逐元素逐通道相加)，形成复合特征图，完成高层语义信息的嵌入；其次，复合特征图和经过下采样的低层特征图相融合，形成混合特征图，完成低层细节信息的嵌入；最后，各混合特征图经过设计的复合卷积层后，生成最终的待检测特征图并进入下一模块。以图 5 中各符号为例，具体的操作步骤为：

结构：LEFPM

输入：C3, C4, P5′

1) P4′经过 1×1 卷积下采样后与 P5′融合生成 x，x 经过复合卷积层后生成 P5；

2) x 经过插值上采样后与 C4 融合，融合后的特征图与 C3(先经过 1×1 卷积层下采样) 相加生成 y，y 经过复合卷积层后生成 P4；

3) y 经过插值上采样后与 C3 融合生成 z，z 经过复合卷积层后生成 P3。

输出：P3, P4, P5

3　实验与分析

3.1　数据集和评价指标

3.1.1　数据集

PASCAL VOC 挑战赛[25] 是一个世界级的计算机视觉比赛，包含目标分类、目标检测、目标分割动作分类等多个子任务。VOC2007 和 VOC2012 是目标检测子任务的两个基准数据集，共包含人、猫、汽车等 20 个类别，每个版本的数据集都采用统一的制作方式，并按照 1:1:2 的比例分别划分为训练集、验证集和测试集，具体的图片和目标数量见表 1。

表 1　VOC 数据集信息

数据集	训练集		验证集		训练+验证集		测试集		总数	
	图片数/张	目标数/个	图片数/张	目标数/个	图片数/张	目标数/个	图片数/张	目标数/个	图片数/张	目标数/个
VOC2007	2501	6301	2510	6307	5011	12608	4952	12032	9963	24640
VOC2012	5717	13609	5823	13841	11540	27450	11540	27450	23080	54900
共计	8218	19910	8333	20148	16551	40058	16492	39482	33043	79540

3.1.2 评价指标 mAP

IoU(intersection-over-union) 为目标预测框和真实框的交集和并集的比值：

$$\mathrm{IoU}=\frac{\mathrm{BBox}_{\mathrm{pred}}\cap\mathrm{BBox}_{\mathrm{gt}}}{\mathrm{BBox}_{\mathrm{pred}}\cup\mathrm{BBox}_{\mathrm{gt}}} \tag{2}$$

若设定 IoU 的阈值为 A，当一个预测框与一个真实框的 IoU 值大于该阈值时，判定为真正例 (true positive, TP)，反之则判定为假正例 (false positive, FP)。

精确率 (precision) 是指预测为正样本的数据中，真正例所占的比重。召回率 (recall) 是指在实际为正样本的数据中，判定为真正例的比重。二者分别作为纵、横坐标组成 P-R 曲线，曲线下的面积称为平均精确率 (average precision, AP)，是对不同召回率点上精确率的积分和。AP 的值越大，说明模型的平均精确率越高：

$$\begin{aligned}\mathrm{Precision}&=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}\\ \mathrm{Recall}&=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}\end{aligned} \tag{3}$$

$$\mathrm{AP}=\int_0^1 P(R)\mathrm{d}(R) \tag{4}$$

一般的，数据集中会包含多种类别，按照类别进行算术平均后的精确率被称为平均精确率均值 (mean average precision, mAP)：

$$\mathrm{mAP}=\frac{1}{\mathrm{classes}}\sum_{i=1}^{\mathrm{classes}}\int_0^1 P(R)\mathrm{d}(R) \tag{5}$$

3.2 损失函数

本文定义训练损失函数如下：

$$L\left(\{p_{x,y}\},\{t_{x,y}\}\right)=\frac{1}{N}\left(\sum_{x,y}L_{\mathrm{cls}}\left(p_{x,y}\right)+\sum_{x,y}kL_{\mathrm{reg}}\left(t_{x,y}\right)\right) \tag{6}$$

式中，L_{cls}使用 Focal Loss 损失函数[26]；L_{reg}使用 GIoU Loss 损失函数[27]；$p_{x,y}$表示类别的预测概率；$t_{x,y}$表示回归的预测坐标；N表示正样本数；k 是指示函数，若确定当前预测为正样本则为 1，否则为 0。

本文算法的推理过程为：给定一张图片，经过模型处理后得到其特征图上每个位置的类别概率$p_{x,y}$和回归坐标$t_{x,y}$，在分类之后选择$p_{x,y}>0.05$的$t_{x,y}$为正样本的坐标，进而确定目标的预测框。

3.3 参数设置

在本实验中，输入到模型的图像大小设为 512×800，在将图像输入到网络之前，对图像数据进行归一化增强处理；为加快收敛，训练时 Res2Net50 加载官方给出的预训练模型；训练优化器采用随机梯度下降策略 (stochastic gradient descent，SGD)[28] 更新网络参数，其中动量 (momentum) 和权值衰减 (weight decay) 参数分别设为 0.9 和 0.000 1；学习率 (learning rate) 的变化采用预热 (warm up) 策略[29]，减缓模型在训练初期对小批量 (mini-batch) 样本的过拟合现象，也有助于保持模型深层的稳定性，初始学习率为 0.01，结束学习率为 0.000 01；后处理 NMS-IoU 的阈值设置为 0.6；在 Qurdro RTX 8000 上单卡 (内存 48 G)训练 30 个周期 (Epochs) 后结束。

3.4 消融实验

本文以 FCOS 为 Baseline，在 VOC2007 和 VOC2012 的训练+验证集上进行训练，在 VOC2007 的测试集上进行验证，分别验证提出模块的有效性。

如表 2 所示，本文通过 3 组递进实验，逐一验证各模块的有效性。首先，只将预训练的 Res2Net50 替换原 Baseline 中的 ResNet50，由第一组实验可知，替换后的模型有 0.7% 的精度提升，而参数量只有很微小的增加，证明了 Res2Net50 在特征提取上的有效性；其次，在第一组实验的基础上，加入本文提出的 HRFM，由第二组实验可知，加入 HRFM 后的模型精度提升明显，由 79.4% 提升到 84.4%，充分证明增强顶层特征图的感受野可以有效解决目标遮挡问题，同时对大尺度目标具有较强的适应性；最后一组实验中，本文将改进后的 LEFPM 取代 Baseline 中的 FPN。改进后的模型精度比第二组实验提高了 2%，相比原 Baseline 在 mAP 上高出 7.6%(如图 6 所示，在各个类别上相较于 Baseline 也有显著提升)，表明低层特征信息对小尺度目标检测非常重要，也为多尺度融合方式提供了参考。另一方面，在引入“Res2Net50”“LEFPM”“HRFM”3 个模块后，本文 ACFPN算法的参数量仅仅比原 Baseline 增加了 0.7 Mb，从侧面反应出引入的特征增强模块并没有增加模型整体的复杂程度和计算量。

表 2 本文提出模块的性能对比

方法	mAP/%	参数量/Mb
ResNet50+FPN (Baseline)	78.7	123.49
Res2Net50+FPN	79.4	124.18
Res2Net50+FPN+HRFM	84.4	125.24
Res2Net50+LEFPM+HRFM (本文 ACFPN)	86.4	124.19

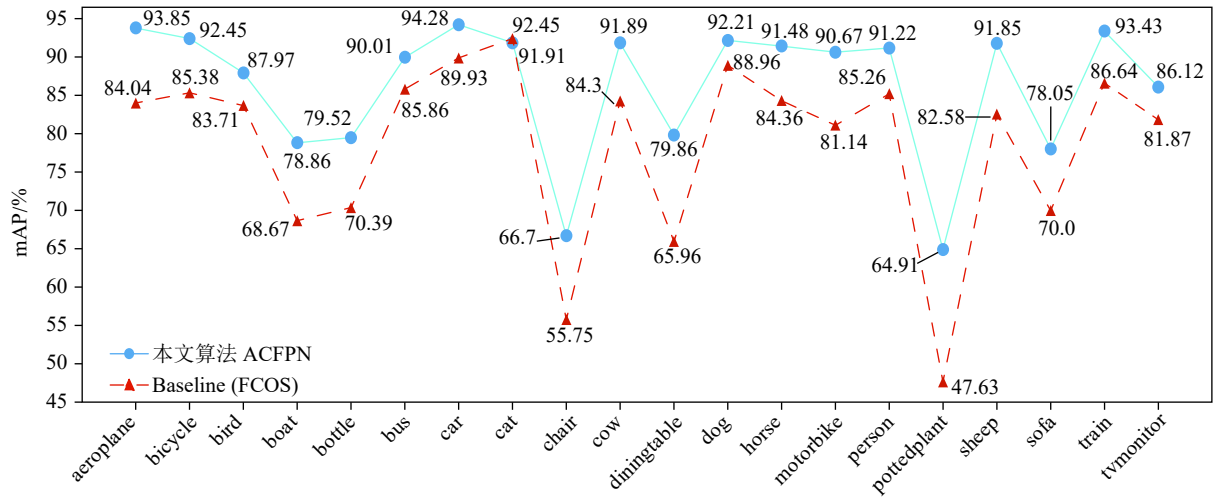


图 6　各类别检测精度对比图

接下来，本文通过对 Loss 进行画图分析，进一步比较所提出方法较原 Baseline 方法的优越性。

由图 7 可以看出，本文的 ACFPN 与原 Baseline 相比，Loss下降迅速，在经过约 80 次迭代之后，Loss 趋于稳定水平。充分证明主干网络 Res2Net50 的引入，有效提高了模型的收敛速度和收敛性。

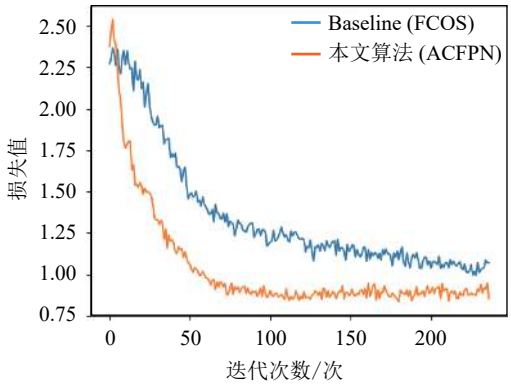


图 7　Loss 曲线图

综合表 2 和图 7 可知，提出的 HRFM 和改进的 LEFPM，得益于对卷积核尺寸和通道数约束的有效设计，在对模型精度提升的同时，并未产生较多的参数量。Res2Net 中分层级联的设计模式，使得在保持提取特征有效性的同时，加快了梯度的反向传播。

3.5　算法对比

为了证明 ACFPN 的整体有效性，本文选取了一系列顶会论文中具有代表性的算法进行对比，结果如表 3 所示，其中 NAS Yolo 是 PASCAL VOC Challenge 榜单上的 Top 1 解决方案[30]。

选取表 3 中对比算法的原则如下：

1) 2018 年以后文献中出现的优秀检测算法；

2) 在权威 VOC 测试集上进行过测试。

表 3　各算法精度对比

时间	算法	数据集	mAP/%
2018	Pelee[31]	VOC07+12	70.9
2018	SIN[32]	VOC07+12	76.0
2019	FCOS[18]	VOC07+12	78.7
2018	HKRM[33]	VOC07+12	78.8
2018	MLKP[34]	VOC07+12	80.6
2018	STDN[35]	VOC07+12	80.9
2019	R-DAD[36]	VOC07+12	81.2
2018	RFBNet[23]	VOC07+12	82.2
2018	RefineDet[37]	VOC07+12	83.8
2018	PFPNet[38]	VOC07+12	84.1
---	**本文算法**	VOC07+12	**86.4**
2020	NAS Yolo (Top 1)[28]	VOC07+12	86.5

由表 3 所示，本文列出了近 3 年间顶会论文中，各方法在 VOC 数据集上的测试结果。可以看出，本文提出的 ACFPN 在精度上超越了各论文方法，比最好的论文方法 PFPNet 在 mAP 上高出 2.3%，距离非论文方法 NAS Yolo(榜单 Top1) 相差仅仅 0.1%，充分证明了本文 ACFPN 方法的有效性和优越性。

3.6　算法效果展示

由检测效果图 8 所示，本文的 ACFPN 方法对比原 Baseline (FCOS) 方法，具有以下明显优势：

1) 定位框的位置更加精准；

2) 召回率更高；

3) 多尺度目标检测、遮挡目标检测效果显著。

a. Baseline方法 (FCOS) b. 本文方法 (ACFPN)

c. Baseline 方法 (FCOS) d. 本文方法 (ACFPN)

e. Baseline 方法 (FCOS) f. 本文方法 (ACFPN)

图 8 检测效果对比图

4 结 束 语

针对目标检测领域普遍存在的遮挡和多尺度问题，本文提出了一种基于空洞卷积特征金字塔的目标检测算法。利用空洞卷积可以有效增大感受野的优点，设计了混合感受野模块 HRFM，采用多种不同尺寸的空洞卷积层密集连接，有效规避了单一空洞卷积造成的网格效应；在现有 FPN 的基础上重新构建网络结构，将低层特征图包含的细节信息嵌入到高层语义信息中，弥补算法对小目标物体的漏检缺陷，进一步提高目标定位的准确率。特别地，在主干部分，ACFPN 将 Res2Net50 代替了常用的 ResNet50，在增强特征表征能力的同时加快了模型收敛速度。Anchor Free 机制可以有效降低候选框的冗余，从而提高定位精度，本文将 FCOS 的这一机制保留。通过在 VOC 数据集上进行测试，本文的 ACFPN 可以达到 86.4% 的 mAP。本文方法为接下来行人重识别任务的开展提供了部分解决思路。

参 考 文 献

[1] 许德刚, 王露, 李凡. 深度学习的典型目标检测算法研究综述[J]. 计算机工程与应用, 2021, 57(8): 10-25.
XU D G, WANG L, LI F. Overview of research on typical target detection algorithms for deep learning[J]. Computer Engineering and Applications, 2021, 57(8): 10-25.

[2] VIOLA P, JONES M J. Robust real-time face detection[J]. International Journal of Computer Vision, 2004, 57(2): 137-154.

[3] DALAL N, TRIGGS B. Histograms of oriented gradients for human detection[C]// 2005 IEEE computer society conference on Computer Vision and Pattern Recognition (CVPR). New York: IEEE, 2005: 886-893.

[4] PAPAGEORGIOU C P, OREN M, POGGIO T. A general framework for object detection[C]// The Sixth International Conference on Computer Vision (ICCV). Bombay: IEEE, 1998: 555-562.

[5] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. Imagenet classification with deep convolutional neural networks[J]. Advances in Neural Information Processing Systems, 2012, 25: 1097-1105.

[6] 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 (CVPR). Columbus: IEEE, 2014: 580-587.

[7] 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.

[8] 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 (CVPR). Las Vegas: IEEE, 2016: 779-788.

[9] REDMON J, FARHADI A. YOLO9000: Better, faster, stronger[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Hawaii: IEEE, 2017: 7263-7271.

[10] Farhadi A, Redmon J. Yolov3: An incremental improvement [EB/OL].[2018-04-08].https://arxiv.org/abs/1804.02767v1.

[11] LIU W, ANGUELOV D, ERHAN D, et al. Ssd: Single shot multibox detector[C]// European Conference on Computer Vision (ECCV). Amsterdam: Springer, 2016: 21-37.

[12] HE K, ZHANG X, REN S, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(9): 1904-1916.

[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 (CVPR). Hawaii: IEEE, 2017: 2117-2125.

[14] YU F, KOLTUN V. Multi-scale context aggregation by dilated convolutions[EB/OL]. [2016-04-30]. https://arxiv.org/abs/1511.07122.

[15] GIRSHICK R. Fast r-cnn[C]// Proceedings of the IEEE

International Conference on Computer Vision (ICCV). Santiago: IEEE, 2015: 1440-1448.
[16] LAW H, DENG J. Cornernet: Detecting objects as paired keypoints[C]// Proceedings of the European Conference on Computer Vision (CVPR). Munich: Springer, 2018: 734-750.
[17] DUAN K, BAI S, XIE L, et al. Centernet: Keypoint triplets for object detection[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Seoul: IEEE, 2019: 6569-6578.
[18] TIAN Z, SHEN C, CHEN H, et al. Fcos: Fully convolutional one-stage object detection[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Seoul: IEEE, 2019: 9627-9636.
[19] GAO S, CHENG M M, ZHAO K, et al. Res2net: A new multi-scale backbone architecture[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019
[20] WANG P, CHEN P, YUAN Y, et al. Understanding convolution for semantic segmentation[C]//2018 IEEE Winter Conference on Applications of Computer Vision (WACV). Nevada: IEEE, 2018: 1451-1460.
[21] SZEGEDY C, LIU W, JIA Y, et al. Going deeper with convolutions[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston: IEEE, 2015: 1-9.
[22] ZHAO H, SHI J, QI X, et al. Pyramid scene parsing network[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Hawaii: IEEE, 2017: 2881-2890.
[23] LIU S, HUANG D. Receptive field block net for accurate and fast object detection[C]//Proceedings of the European Conference on Computer Vision (ECCV). Munich: Springer, 2018: 385-400.
[24] NAJIBI M, SAMANGOUEI P, CHELLAPPA R, et al. Ssh: Single stage headless face detector[C]//Proceedings of the IEEE International Conference on Computer Vision (ICCV). Venice: IEEE, 2017: 4875-4884.
[25] EVERINGHAM M, VAN GOOL L, WILLIAMS C K I, et al. The pascal visual object classes (voc) challenge[J]. International Journal of Computer Vision, 2010, 88(2): 303-338.
[26] 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 (ICCV). Venice: IEEE, 2017: 2980-2988.
[27] REZATOFIGHI H, TSOI N, GWAK J Y, et al. Generalized intersection over union: A metric and a loss for bounding box regression[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach: IEEE, 2019: 658-666.
[28] BOTTOU L. Stochastic gradient descent tricks[M]// Neural networks: Tricks of the trade. Berlin: Springer, 2012: 421-436.
[29] FAN X, JIANG W, LUO H, et al. Spherereid: Deep hypersphere manifold embedding for person re-identification[J]. Journal of Visual Communication and Image Representation, 2019, 60: 51-58.
[30] EVERINGHAM M, VAN GOOL L, WILLIAMS C K I, et al.PASCAL VOC Challenge performance evaluation and download server[EB/OL]. [2021-1-30]. http://host.robots.ox.ac.uk:8080/leaderboard/displaylb_main.php?challengeid=11&compid=3.
[31] WANG R J, LI X, LING C X. Pelee: A real-time object detection system on mobile devices[EB/OL]. [2019-1-18]. https://arxiv.org/abs/1804.06882.
[32] LIU Y, WANG R, SHAN S, et al. Structure inference net: Object detection using scene-level context and instance-level relationships[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City: IEEE, 2018: 6985-6994.
[33] JIANG C, XU H, LIANG X, et al. Hybrid knowledge routed modules for large-scale object detection[EB/OL]. [2018-10-30]. https://arxiv.org/abs/1810.12681.
[34] WANG H, WANG Q, GAO M, et al. Multi-scale location-aware kernel representation for object detection[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City: IEEE, 2018: 1248-1257.
[35] ZHOU P, NI B, GENG C, et al. Scale-transferrable object detection[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City: IEEE, 2018: 528-537.
[36] BAE S H. Object detection based on region decomposition and assembly[C]//Proceedings of the AAAI Conference on Artificial Intelligence (AAAI). Hawaii: AAAI, 2019, 33(1): 8094-8101.
[37] ZHANG S, WEN L, BIAN X, et al. Single-shot refinement neural network for object detection[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City: IEEE, 2018: 4203-4212.
[38] KIM S W, KOOK H K, SUN J Y, et al. Parallel feature pyramid network for object detection[C]//Proceedings of the European Conference on Computer Vision (ECCV). Munich: Springer, 2018: 234-250.

编　辑　刘飞阳