计算机科学与探索
Journal of Frontiers of Computer Science and Technology
1673-9418/2023/17(05)-1102-10
doi: 10.3778/j.issn.1673-9418.2109099

引入残差学习与多尺度特征增强的目标检测器

贾天豪[1],彭　力[1+],戴菲菲[2]
1. 物联网技术应用教育部工程研究中心(江南大学 物联网工程学院),江苏 无锡 214122
2. 台州市质量安全检测研究院,浙江 台州 318020
+ 通信作者 E-mail: penglimail2002@163.com

摘　要:目前深度学习在计算机视觉领域中取得了巨大成功,但是小目标检测仍是目标检测领域中具有挑战性的难题。针对小物体分辨率低、图像模糊、携带信息少等问题,提出了引入残差学习与多尺度特征增强的目标检测器。首先在主干网络中引入基于残差学习的增强特征映射块,通过通道平均和归一化处理使得模型更加专注于对象区域而不是背景,并在兼顾检测速度的同时为有效特征层提供额外的语义信息;然后特征映射对上下文信息敏感的特征融合块进一步增大有效特征图的感受野,并将用于预测的浅特征层与深特征层进行融合,提高低分辨率下的检测性能;最后通过双重注意力块抑制背景噪音,将关键特征嵌入到注意力中,在保留空间信息的同时加强通道间的信息关联,进而增强特征的表达能力。为了更好地检测小目标,还对浅层特征映射先验框数量进行了调整。实验结果表明,在PASCAL VOC2007的数据集上,该算法对于300×300输入尺度的检测精度(mAP)为79.9%,较SSD提高了2.7个百分点,对小目标bird、bottle、chair、plant检测精度分别提升了5.1个百分点、7.5个百分点、3.9个百分点、7.2个百分点。在OAP自制航拍数据集上的检测精度(mAP)为82.7%。

关键词:目标检测;残差学习;卷积神经网络(CNN);注意力机制
文献标志码:A　**中图分类号:**TP391.4

Object Detector with Residual Learning and Multi-scale Feature Enhancement

JIA Tianhao[1], PENG Li[1+], DAI Feifei[2]
1. Engineering Research Center of Internet of Things Technology Applications (School of Internet of Things Engineering, Jiangnan University), Ministry of Education, Wuxi, Jiangsu 214122, China
2. Taizhou Institute of Quality and Safety Testing, Taizhou, Zhejiang 318020, China

Abstract: At present, deep learning has achieved great success in the field of computer vision, but small object detection is still a challenging problem in the field of object detection. Aiming at the problems of low resolution of small objects, blurred images, and less information carried, one object detector that introduces residual learning and multi-scale feature enhancement is proposed. Firstly, an enhanced feature mapping block based on residual learning is introduced into the backbone network. Through channel averaging and normalization, the model more focuses on the object area instead of the background, and it provides additional semantics information for the effective feature layer while taking into account the detection speed. Then the feature map increases the receptive field of the effective feature map through feature fusion block sensitive to context information, and fuses the shallow feature layer and the deep feature layer used for prediction to improve the detection performance at low resolution. Finally,

基金项目:国家重点研发计划(2018YFD0400902);国家自然科学基金(61873112)。
This work was supported by the National Key Research and Development Program of China (2018YFD0400902), and the National Natural Science Foundation of China (61873112).
收稿日期:2021-09-27　**修回日期:**2021-11-15

a dual attention block is used to suppress background noise, and key features are embedded in attention. While preserving spatial information, it strengthens the information association between channels, thereby enhancing the expressive ability of features. In order to better detect small objects, the number of a priori boxes for shallow feature mapping is also adjusted. Experimental results show that on the dataset of PASCAL VOC2007, the detection accuracy (mAP) of the algorithm for 300×300 input scale is 79.9%, which is 2.7 percentage points higher than that of SSD, and the detection accuracy of small objects bird, bottle, chair, and plant is improved 5.1 percentage points, 7.5 percentage points, 3.9 percentage points, 7.2 percentage points, respectively. The detection accuracy (mAP) on the OAP self-made aerial dataset is 82.7%.

Key words: object detection; residual learning; convolutional neural network (CNN); attention mechanism

随着人工智能技术的飞速发展、深度卷积网络的出现[1]，引入了一些能学习语义、高水平、深层次特征的工具来解决传统体系结构中存在的问题，使模型在网络架构、训练策略和优化功能方面的性能得到了显著提高[2-4]。然而，图像中小尺度目标区域相对较小、图像模糊、信息量不足，导致在卷积神经网络模型中对多尺度、低分辨率、小目标检测的研究一直是个难题。

目前，视觉任务中解决该问题主要分为两个方向：一是使用图像金字塔[5-6]的方式，对图像进行一定比例的缩放，从而得到一系列不同尺寸的样本图像序列，在缩放过程中采用线性差值等方法进行上采样，同时还可以加入滤波、模糊等处理方式丰富样本的细节信息。二是使用特征金字塔[7-8]的方式，通过利用常规卷积神经网络（convolutional neural network，CNN）模型内部从底至上各个层对同一尺度图片不同维度的特征表达结构，在单一图片视图下生成对其的多维度特征表达，可以有效地赋能CNN模型，从而生成表达能力更强的特征图。

其中，Liu等人提出的一阶段目标检测器（single shot multibox detector，SSD）[9]在基础网络的顶端额外增加了更多卷积层来构成特征金字塔，并利用不同层次的特征图定义预选框进行最终预测，这种策略使得小目标在浅层不会丢失太多的位置信息，而大目标在深层也可以很好地定位和识别。Li等人提出利用特征金字塔网络融合模型高低层语义信息，并将融合结果用于生成新的特征金字塔，从而增强小目标的特征表达能力（feature fusion single shot multibox detector，FSSD）[10]。Singh等人从训练角度切入，在数据的层面思考，采用了一种多尺度的训练方式——图像的尺度归一化（analysis of scale invariance in object detection，SNIP）[11]，在金字塔模型的每一个尺度上进行训练，高效利用训练数据，检测效果得到显著提升。Fu等人提出通过更换主干网络并加入反卷积的方式来降低小目标漏检率的目标检测算法（deconvolutional single shot detector，DSSD）[12]。Zhou等人提出scale-transfer Module对特征图进行放大或缩小，并分别对不同尺度的特征图做目标预测（scale-transferrable object detection，STDN）[13]。宋云博等人提出了基于级联卷积神经网络的高效目标检测算法[14]。

虽然上述采用特征金字塔结构的目标检测器拥有不错的检测效果，但是它们在处理多尺度问题时没有考虑到全局的上下文信息对小目标检测的影响，并且没有进一步增强不同尺度下的关键特征信息，这使得小目标和低分辨率目标检测性能还有进一步提升的空间。

因此，本文在SSD[9]算法基础上提出了引入残差学习与多尺度特征增强的目标检测器（object detector with residual learning and multi-scale feature enhancement，RMFE-SSD）。在特征提取网络中加入增强特征映射块（enhanced feature map block，EFB）来丰富有效特征层的语义信息，使得模型更加专注于对象区域。构建含有上下文信息的特征融合块（context-sensitive feature fusion module，CFB），在增大特征感受野的同时，将细节信息丰富的浅层特征与语义信息丰富的深层特征进行融合，并通过双重注意力块（double attention block，DAB）将关键特征嵌入到注意力中，实现空间与通道间的信息关联，从而增强模型的特征表达能力，提高模型的目标检测性能。

1 引入残差学习与多尺度特征增强的目标检测器

本文提出的RMFE-SSD模型包含7个增强特征映射块、2个携带上下文的特征融合块和2个双重注意力块，主干网络仍采用VGG16[15]进行特征提取。其中，7个EFB用于提高SSD目标检测器中有效特征

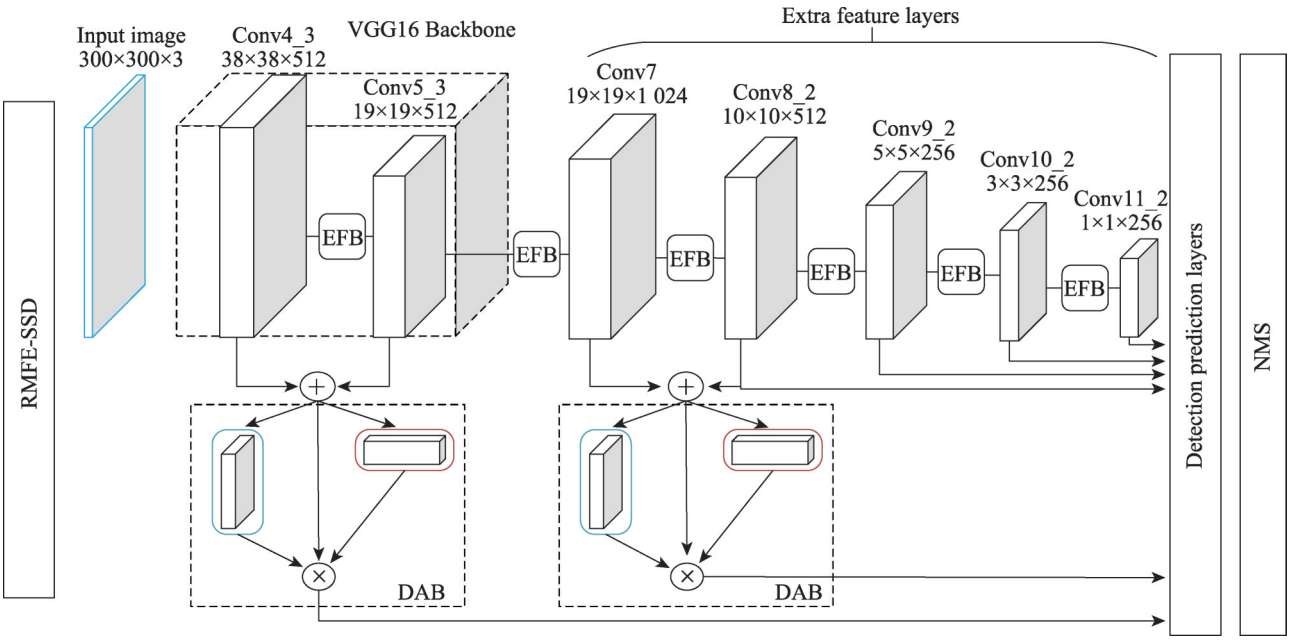

图1 整体网络结构图

Fig.1 Overall network structure diagram

图的特征表达能力，专注于学习除背景外的目标区域。采用2个CFB来扩大主干网络中浅层特征的感受野，并对Conv4_3、Conv5_3和Conv7、Conv8_2两组不同尺度特征进行融合操作，使得浅层的特征图具有更多的语义信息，能够更好地捕获小目标。2个DAB用于突出关键特征，加强特征在空间和通道上的信息关联。整体架构如图1所示。

1.1 增强特征映射块

在基于卷积神经网络的目标检测模型中，研究者们普遍认为网络深度越深，模型的非线性的表达能力越强，学习效果越好。然而，网络的不断加深会导致模型退化、错误率升高的问题。对此，He等人[16]提出在残差网络中引入跳跃连接，将一个潜在的恒等映射转换为对残差函数的学习，可以有效地去除相同的特征主体，体现它们之间的差异性，如图2(a)所示。因此，本文以ResNet[17]中提出的残差块为基础来构建增强特征映射块，如图2(b)所示。通过反向传播的不断学习，可以有效增强特征图的细节特征信息。

本文所提出的增强特征映射块与ResNet[14]中的残差块相比存在两点优势：(1)本文采用局部特征学习来抑制卷积操作带来的计算量，兼顾检测速度与检测精度。将输入特征映射 f 分割为 $f_{1/4}\in\mathbf{R}^{H\times W\times\frac{C}{4}}$ 和 $f_{3/4}\in\mathbf{R}^{H\times W\times\frac{3}{4}C}$ 两部分，$f_{3/4}$ 通过3个卷积层提取更多的语义特征后与 $f_{1/4}$ 进行concat连接，其中第一个卷积操作采用1×1的卷积核对特征图进行降维，第二个卷积操作采用3×3的卷积核进行特征提取，第三个卷积操作采用1×1的卷积核进行升维。(2)本文对输入特征映射 f 额外采用通道平均池化和归一化处理来提高特征的可辨别性，并在全局范围内捕获更有

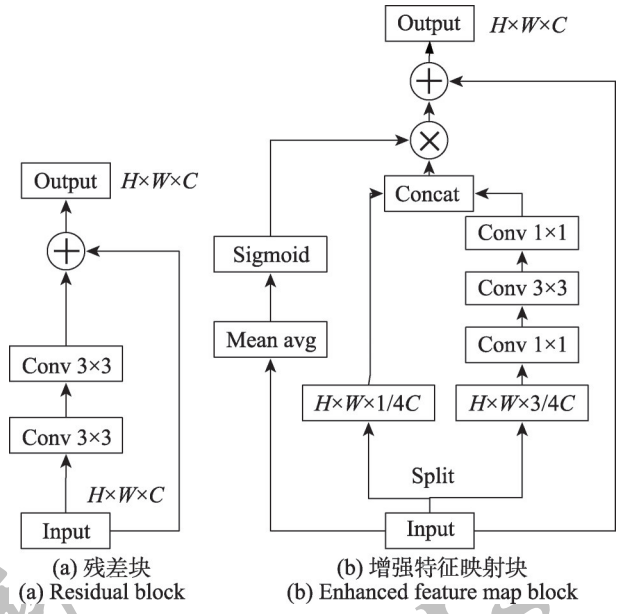

图2 残差块与增强特征映射块结构图

Fig.2 Structure diagram of residual block and enhanced feature map block

效的细节信息，使得图像的语义性愈加强烈，进而增强了用于预测的特征金字塔的表达能力。

卷积神经网络能够提取low/mid/high层的特征，网络的层数越多，意味着网络提取的特征越抽象，越具有语义信息。本文设计的EFB模块通过引入残差块来加深网络并融合上述两点优势，在兼顾检测速度的同时通过归一化处理突出了附加在通道内的语义信息，提高了模型的检测性能。该模块的输出结果通过式(1)进行计算。

$$f_o = \sigma(A(f)) \times O(C_2^{1\times1}(C_1^{3\times3}(C_0^{1\times1}(f_{3/4}))), f_{1/4}) \quad (1)$$

其中，C 表示卷积操作，上标表示卷积核大小，O 表示融合操作，σ 表示sigmoid，A 表示平均池化。

此外，考虑到精度和速度之间的权衡，本文通过实验验证了将输入特征映射分割成3/4和1/4是最有效的方法（详见2.3节）。按照这种方式，该模块可以在兼顾检测速度的同时提升小目标检测精度。

1.2 对上下文信息敏感的特征融合块

现有的特征融合模块都是将不同尺度的特征图直接通过上采样后进行融合，这忽略了部分关键的上下文信息，尤其对小目标和低分辨率特征图而言。本文提出先采用对上下文信息敏感的特征块CFB扩大检测网络的感受野范围，丰富对象区域的上下文信息，然后进行基于反卷积的上采样融合操作。由于concat操作只是在通道维度上将不同尺度的特征连接，不能反映不同通道间特征的相关性和重要性[18]，本文算法中融合方式采用的是element-sum。

CFB模块的结构如图3所示，其主要包含3个分支，并在不同分支上设定不同大小的步长和卷积核，使得网络具有了更宽的特征映射块。其中cfb1和cfb2两个分支分别通过3×3的卷积核削减通道数量，获取全局的特征信息；cfb3和cfb4子分支在cfb2主分支下通过不同大小的卷积核捕获不同感受野下的上下文信息。特别地，本文将5×5的卷积核替换为2个3×3的卷积核，一方面通过堆叠3×3卷积核提供了更多数量的激活函数，增加网络的非线性；另一方面是卷积操作本身并没有破坏图像的空间信息，大感受野不具有优势，反而会增加计算量，本文在2.2节进行了对比验证。对于输入in、输出out的计算公式为：

$$f_{\text{out}} = O(f_{\text{res}}, C_0^{3\times3}(f_{\text{res1}}), C_1^{3\times3}C_2^{3\times3}(f_{\text{res1}})) \quad (2)$$

其中，C 表示卷积操作，上标表示卷积核大小，O 表示融合操作，$f_{\text{res}} = C_1^{3\times3}(\text{in})$，$f_{\text{res1}} = C_2^{3\times3}(\text{in})$。

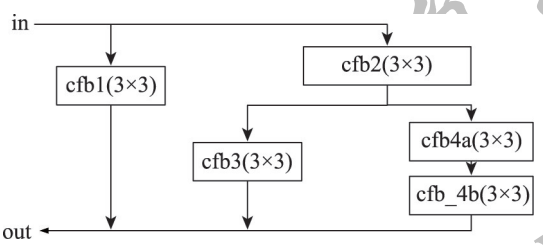

图3 对上下文敏感的特征融合块

Fig.3 Context-sensitive feature fusion module

由于Conv8_2以后的特征图分辨率过低，对融合效果帮助不大，其带来额外的计算量会降低检测效率。本文仅选择主干网络中的Conv4_3、Conv5_3和额外卷积层中的Conv7、Conv8_2进行特征融合。Conv5_3与Conv7二者拥有同样的分辨率，但是拥有不同的语义信息，选择Conv5_3与Conv4_3进行融合对预测模块来说更有利。如图1所示，Conv4_3特征图的尺寸为38×38×512，Conv5_3特征图的尺寸为19×19×512，二者分别通过CFB模块后进行降维操作，由于浅特征层的特征分布和深特征层之间存在较大的间隙，直接融合效果不好，添加Batch-Norm层进行归一化处理，这样也可加速训练速度，防止梯度消失。Conv5_3再通过反卷积上采样后与Conv4_3进行element-sum操作，融合过程如图4所示。Conv7和Conv8_2的融合过程类似，如图5所示。

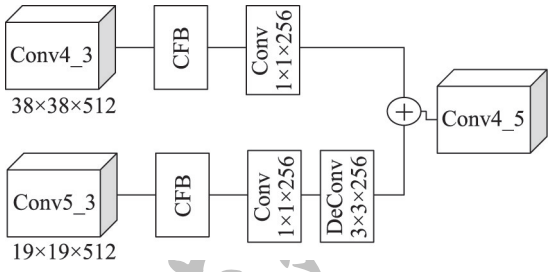

图4 Conv4_3与Conv5_3的融合过程

Fig.4 Fusion process of Conv4_3 and Conv5_3

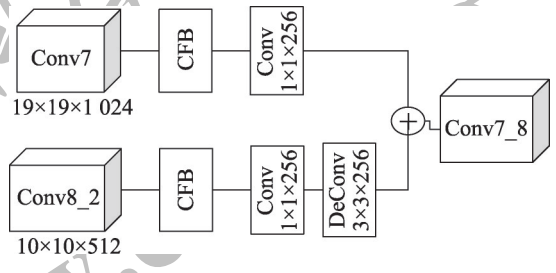

图5 Conv7与Conv8_2的融合过程

Fig.5 Fusion process of Conv7 and Conv8_2

此外，本文对不同融合方式下的检测效果进行了对比（见2.2节），实验表明，element-wise-sum方式比concat拥有更好的效果。

为了验证引入CFB模块对于提取目标特征信息的有效性，本文使用热力图可视化方法来直观地对比添加该模块前后模型对目标区域敏感程度的情况。如图6所示，实验对Conv4_3和Conv5_3两个有效特征图进行融合操作，并对比了使用CFB前后的热力图，图中红色部分越深说明对这部分的关注度越高。从图中可以看出，使用CFB模块后模型对目标区域的关注度更加全面，效果更好，这是因为在深浅特征图的融合操作中，CFB模块扩大了特征图的感受野，丰富对象区域的上下文信息，使得模型可以更加准确地感知学习，该实验也进一步验证了CFB模块的有效性。

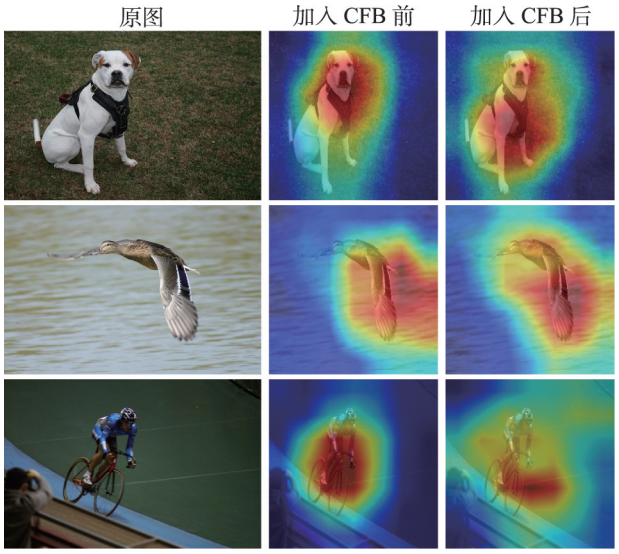

图6 热力图可视化

Fig.6 Visualization of heat maps

1.3 双重注意力块

由于特征图不断被卷积操作压缩，小目标的有效信息变得更少，甚至会被背景信息所覆盖。本文设计了基于ECA-Net(efficient channel attention networks)[19]的双重注意力块，通过空间注意力与通道注意力并联的方式，有效捕获小目标，同时抑制背景信息。

如图7所示，DAB包含空间注意力和通道注意力两部分，空间注意力使用两层感知机进行非线性的特征变换，并利用Sigmoid函数实现特征重标定，为每个位置生成权重掩膜并加权输出，从而使得模型更加聚焦于前景特征而不是背景区域。通道注意力利用全局平均池化和卷积权重共享的方式赋予每个通道不同的权重系数，并自适应地调整通道间的特征响应，区分出重要与非重要的特征信息，使得模型有效地捕获目标区域。

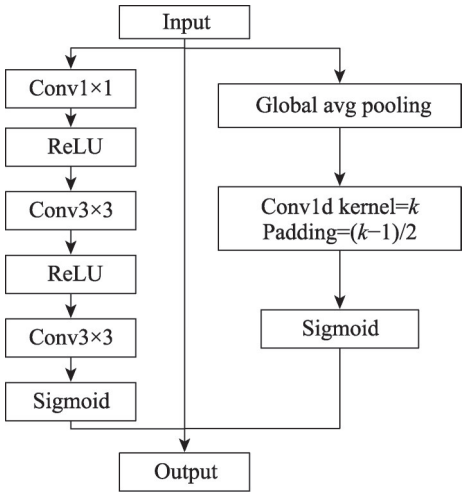

图7 双重注意力块

Fig.7 Double attention block

空间注意力首先通过1×1的卷积核削减特征图的通道数，减少计算量。然后通过两个3×3的卷积核提取空间信息，期间使用ReLU函数进行激活，增加模型的非线性。最后通过Sigmoid函数获取二维空间特征映射，用于对原有特征图的加权。之所以使用3×3的卷积核来提取空间信息，是因为它能够在保证相同感受野的情况下减少参数量。空间注意力输出结果 f_1 的计算过程如式(3)所示。

$$f_1=\sigma(C_3^{3\times3}(C_2^{3\times3}(C_1^{1\times1}(F)))) \tag{3}$$

其中，C 表示卷积操作，上标表示卷积核大小，F 表示原有特征图，σ 表示Sigmoid激活函数。

受ECA-Net[19]的启发，本文的通道注意力模块去除了SENet(squeeze-and-excitation networks)[20]模块中的FC(fully connected)层，直接在全局平均池化(global average pooling，GAP)之后的特征图上通过一个可以权重共享的1D卷积进行学习，并采用自适应选择一维卷积核大小 k 的方法，确定局部跨信道交互的覆盖率，从而实现通道间的信息交互，通道注意力的输出 f_2 的计算过程如式(4)所示。对于不同的通道数 C，超参数 k 拥有不同大小的值，k 和 C 的对应关系如式(5)所示，其中2的次方考虑的是通道数量一般是2的指数倍。

$$f_2=\sigma(\mathrm{C1D}_k(F)) \tag{4}$$

$$C=\Phi(k)=2^{2k-1} \tag{5}$$

其中，C1D表示一维卷积，下标 k 表示卷积核大小，F 表示原特征图，σ 表示Sigmoid激活函数。

综上，本文在SSD[9]原特征金字塔上加入EFB、CFB、DAB三部分模块形成新的特征金字塔，新特征

金字塔弥补了原特征金子塔对小目标有效特征信息丢失、语义信息与细节信息没有充分融合的不足，利用特征重标定、归一化处理和丰富对象区域上下文信息等方法对图像特征进行了增强，可以自适应地调整通道间的特征响应，并区分出重要与非重要的特征信息，这对于模型的检测性能很有帮助。与原特征金字塔相比，展现出了更有效的特征提取手段和更全面的特征表达。

表1展示了图像输入尺度为300×300情况下，用于预测特征金字塔的各层结构参数。在参数的选取过程中，按照如下原则进行选取：(1)通常在达到相同感受野的情况下，卷积核越小，所需要的参数和计算量越小，并且大小为偶数的卷积核即使对称地加padding也不能保证输入特征图尺寸和输出特征图尺寸不变，因此本文选用3×3的卷积核进行特征提取，选用1×1的卷积核进行特征降维或升维。(2)由于目前关于每层卷积的通道数如何选取没有太多的理论支撑，还是根据经验进行设定并通过实验进行验证调整。

表1 用于预测的特征金字塔各层结构参数

Table 1 Structural parameters of each layer of feature pyramid used for prediction

Level	输入	通道数	卷积核	卷积核数	通道数	输出
输入	300×300	3	—	—	—	—
Conv4_3	38×38	512	3×3	512	512	38×38
Conv5_3	19×19	512	3×3	512	512	19×19
Conv4_5_3	38×38 19×19	512 512	1×1 3×3	256	256	38×38
Conv7	19×19	512	3×3	1 024	1 024	19×19
Conv8_2	19×19	512	3×3	512	512	10×10
Conv7_8_2	19×19 10×10	1 024 512	1×1 3×3	256	256	19×19
Conv4_6	38×38	256	3×3	256	256	38×38
Conv7_9	19×19	256	3×3	256	256	19×19
Conv9_2	10×10	512	3×3	256	256	5×5
Conv10_2	5×5	256	3×3	256	256	3×3
Conv11_2	3×3	256	3×3	256	256	1×1

2 实验结果与分析

本文算法是基于深度学习框架Pytorch1.0实现的，计算机操作系统为64位的Ubuntu16.04，内存16 GB，处理器为英特尔i5-8500@3.00 GHz六核，显卡为英伟达GTX 1080Ti，显存11 GB。采用PASCAL VOC公共数据集和自制OAP航拍数据集对算法的有效性进行验证，并分别在300×300和512×512分辨率下，对不同算法的检测性能进行对比。

对比检测算法包括：(1)SSD[9]。(2)一阶段目标检测器，以SSD为基础改进的反卷积单步骤探测器DSSD[12]；基于多层特征做预测，并对预测结果做融合得到最终结果的目标检测器STDN[13]。(3)两阶段目标检测器，一种利用感兴趣区域内部、外部信息的目标检测器(inside-outside net，ION[21])；基于边框回归的实时目标检测器Faster R-CNN[22]。(4)注意力机制对比算法。ECA-Net[19]、SE[20]、在原有通道注意力的基础上衔接空间注意力模块CBAM[23](convolutional block attention module)、移动网络注意力机制Coordinate Attention[24]。

2.1 性能分析

实验1 在PASCAL VOC数据集上的性能对比

PASCAL VOC挑战赛[25]是视觉对象的分类识别和检测的一个基准测试，提供用于训练模型的训练集和评估模型的测试集。该数据集包含vehicle、household、animal、person 4个大类，总共20个小类(加背景21类)。实验在VOC2007和VOC2012的train+val(16 551张)上进行训练，使用VOC2007的test(4 952张)进行测试。训练过程中，初始学习率为0.000 35，300×300分辨率下的batch size设置为32，最大迭代次数设置为120 000，前500个iteration学习率会逐渐增长，该操作可以加速模型的收敛。当iteration是60 000、80 000、100 000时，学习率分别乘以0.1。512×512分辨率下batch size设置为16，最大迭代次数设置为160 000，当iteration是80 000、120 000、140 000时，学习率分别乘以0.1。此外，6个用于预测的特征层所对应的先验框数量分别为6、6、6、6、4、4。对于300×300输入的网络模型的精度变化曲线如图8所示。

表2展示了在PASCAL VOC数据集下算法性能对比，采用AP和mAP作为评估指标，对于最高单类别目标AP进行了加粗显示。对于输入尺寸为300×300时，RMFE-SSD的平均检测精度为79.9%，比SSD算法高2.7个百分点，单类别AP有12项最高，同时也高于DSSD、STDN在内的一阶段目标检测模型，相比ION、Faster R-CNN分别提高了4.3个百分点、6.7个百分点。特别是对bottle、chair、plant等小目标类别上展现了绝对的优势。当输入尺寸为512×512时，RMFE-SSD的平均检测精度为81.7%，与SSD512相比提高了2.2个百分点，同时高于DSSD在内的其

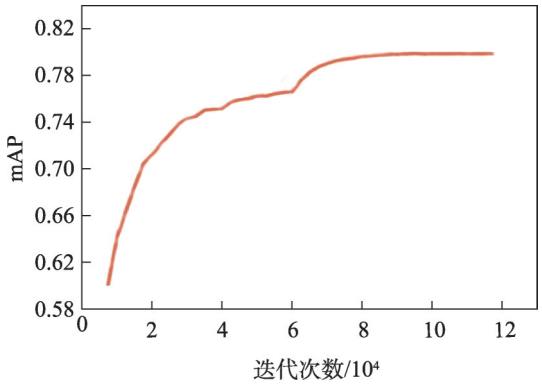

图8 mAP变化曲线

Fig.8 mAP change curve

他一阶段检测算法。在20个类别当中有9个类别的AP最优。这表明了本文模型在不同分辨率下对不同尺度的物体检测性能提升是有效的，在保证检测速度的同时减少了漏检率、误检率，提高了检测精度。

实验2 在OAP自制航拍数据集上的性能对比

OAP（object aerial photography）自制航拍小目标数据集是来自不同传感器和采集平台的航拍样本，由于拍摄距离较远，图像中多以小目标为主。其中包含22 761张来自不同传感器和采集平台的航拍样本，包含了车辆、船舶、飞机等13类（加背景14类）小尺度目标。与VOC数据集相比，目标数量更多、尺寸更小。实验中使用训练集（含有10 818张图片）进行训练，使用测试集（含有10 943张图片）进行测试。训练过程中，在300×300分辨率下最大迭代次数设置为120 000，学习率在80 000和100 000时进行调整，其他参数基本与VOC数据集下的训练参数设置相同。

实验中，除了上述实验的对比算法外，本文还加入了在该数据集下具有不错检测效果的RSSD[26]（rainbow single shot detector）、YOLOv3（you only look once version3）[27]、R-FCN[28]（region-based fully convolutional networks）等算法。表3展示了在OAP航拍数据集上的算法性能对比，同样采用AP和mAP

表2 PASCAL VOC数据集上算法性能对比

Table 2 Performance comparison of algorithms on PASCAL VOC dataset

类别	输入尺寸为300×300下AP值/%						输入尺寸为512×512下AP值/%			
	SSD	DSSD	STDN	ION	Faster R-CNN	RMFE-SSD(ours)	SSD	DSSD	STDN	RMFE-SSD(ours)
Aero	**83.4**	81.9	81.2	79.2	76.5	83.3	84.8	86.6	86.1	**87.1**
Bike	85.2	84.9	88.3	83.1	79.0	**86.2**	85.1	86.2	**89.3**	88.0
Bird	75.0	**80.5**	78.1	77.6	70.9	80.1	81.5	**82.6**	79.5	81.9
Boat	71.2	68.4	72.2	65.6	66.5	**74.1**	73.0	74.9	74.3	**77.7**
Bottle	50.8	53.9	54.3	54.9	53.1	**58.3**	57.8	62.5	61.9	**65.3**
Bus	85.1	85.6	**87.6**	85.4	83.1	86.9	87.8	**89.0**	88.5	88.6
Car	86.1	86.2	86.5	85.1	84.7	**87.2**	88.3	88.7	88.3	**89.1**
Cat	87.0	**88.9**	88.8	87.0	86.4	88.1	87.4	88.8	**89.4**	88.8
Chair	61.4	61.1	63.5	54.4	52.0	**65.3**	63.5	65.2	67.4	**68.3**
Cow	80.9	83.5	83.2	80.6	81.9	**83.9**	85.4	**87.0**	86.5	85.6
Table	76.5	78.1	79.4	73.8	65.7	**80.5**	73.2	**78.7**	79.5	76.7
Dog	84.1	86.7	86.1	85.3	84.8	**87.1**	86.2	**88.2**	86.4	86.6
Horse	87.1	88.7	**89.3**	82.2	84.6	87.5	86.7	89.0	**89.2**	88.1
Mbike	83.6	86.7	**88.0**	82.2	77.5	86.0	83.9	87.5	**88.5**	86.8
Person	78.3	79.7	77.0	74.4	76.7	**81.1**	82.5	83.7	79.3	**84.0**
Plant	47.8	51.7	52.5	47.1	38.8	**55.0**	55.6	51.1	53.0	**57.7**
Sheep	73.5	78.0	80.3	75.8	73.6	**81.0**	81.7	**86.3**	77.9	83.7
Sofa	77.1	**80.9**	80.8	72.7	73.9	79.6	79.0	81.6	81.4	**82.2**
Train	83.2	87.2	86.3	84.2	83.0	**88.0**	86.6	85.7	86.6	**88.0**
TV	76.1	**79.4**	82.1	72.6	72.6	79.2	80.0	83.7	**85.5**	80.5
GPU	Titan X	Titan X	Titan XP	—	Titan X	1080Ti	Titan X	Titan X	Titan XP	1080Ti
FPS	46.0	9.5	40.1	—	7.0	52.0	19.0	5.5	28.6	32.0
mAP/%	77.2	78.6	79.3	75.6	73.2	**79.9**	79.5	81.5	80.9	**81.7**

表3 OAP航拍数据集上算法性能对比

Table 3 Performance comparison of algorithms on OAP aerial photography dataset

Model	AP/%				FPS	mAP/%
	Airplane	Ship	Storage tank	Tennis court		
SSD	79.5	81.9	75.2	69.4	62	78.1
DSSD	81.9	84.9	78.4	70.5	13	79.5
RSSD	80.7	83.2	77.1	69.8	35	78.7
R-FCN	76.6	80.3	74.2	68.5	27	77.1
Faster R-CNN	74.3	78.7	71.9	64.5	11	72.5
YOLOv3	86.2	**85.7**	77.3	74.6	66	82.1
RMFE-SSD	**90.8**	84.5	**89.6**	**87.7**	50	**82.7**

作为评估指标，对于最高单类别目标AP进行了加粗显示。从实验结果可以看出，RMFE-SSD的平均检测精度为82.7%，比SSD提高了4.6个百分点。特别是对于Airplane、Ship、Storage tank和Tennis court等小目标，本文算法拥有很大的精度提升，高于其他检测算法，具有绝对优势。

图9展示了在VOC和OAP航拍数据集上对SSD和RMFE-SSD检测算法的检测结果对比，每组对比图中位于上位置或左侧位置的为原始SSD检测结果。从图中可以看出，SSD检测算法的检测结果存在检测不出和漏检的情况，对远处目标和密集小目标的检测效果不好，RMFE-SSD算法可以很好地处理这些问题。因此，综合上述实验表明，RMFE-SSD检测模型拥有更好的检测性能。

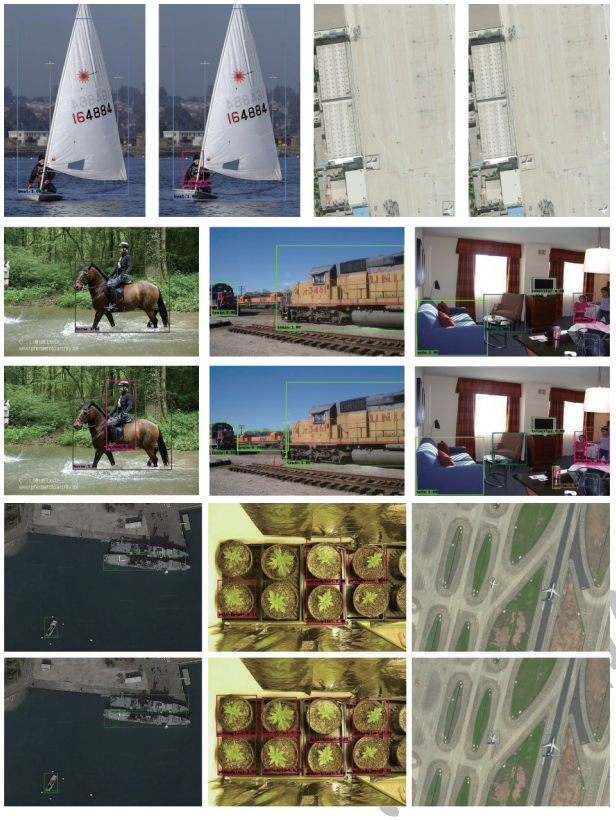

图9 SSD和RMFE-SSD在VOC和OAP数据集上的对比

Fig.9 Comparison of SSD and RMFE-SSD on VOC and OAP datasets

2.2 消融实验

为了进一步评估RMFE-SSD目标检测模型中不同模块的有效性，本文分别对增强特征映射块、对上下文敏感的特征融合模块、双重注意力模块等进行了消融实验研究。所有实验均是使用PASCAL VOC 2007和PASCAL VOC2012训练集进行训练，使用PASCAL VOC2007测试集进行测试。表4展示了使用不同模块下的检测结果。

表4 各模块有效性对比

Table 4 Comparison of effectiveness of each module

EFB	DAB	CFB	Fusion method	Num_priors (38×38)	mAP/%
√			element-sum	6	78.5
√		2×3	element-sum	6	79.4
√	√	2×3	element-sum	4	79.3
√	√	2×3	concat	6	79.6
√	√	1×5	element-sum	6	79.6
√	√	2×3	element-sum	6	**79.9**

从表4的第1行可以看出，加入EFB模块可以使模型的检测精度提高1.3个百分点；从第1行和第2行对比看出，在EFB模块的基础上，加入CFB模块使检测精度再次提高0.9个百分点，这说明融入上下文信息的特征融合模块可以有效提高模型的检测性能；从第2行和第6行可以看出，DAB模块的加入使模型精度再度提高0.5个百分点。不同的融合方式，模型的检测性能也是不同的，从第4行和第6行对比看出，采用对应元素相加的融合方式比直接串联拥有更好的检测效果。从第3行和第6行对比看出，将第一个用于预测的特征图的先验框数量调整为6后，模型的检测性能有所提高。此外，从第5行和第6行对比看出，将5×5的卷积核替换为两个3×3的卷积核，不仅可以提高检测的实时性，还能提高模型的检测精度。综上所述，本文所提出的相关模块对模型的检测性能均起到了积极作用。

2.3 不同分割比例对比实验

在EFB模块中，原始特征图被分为两部分后进行特征语义的提取。为了比较不同分割比例对模型

检测精度的影响，将分割比例设置为1/4、2/4、3/4、4/4。在实验中，使用PASCAL VOC2007测试集测试模型的检测性能。

实验结果如表5所示，从表5中第3行和第4行可以看出，这两种不同分割比例下检测精度相同，但是前者的检测速度快于后者，因此本文算法也采用第3行所示的比例进行分割。

表5 分割比例对模型的影响

Table 5 Effect of split ratio on model

Method	1/4	2/4	3/4	4/4	FPS	mAP/%
EFB_1/4	√				55	79.3
EFB_2/4		√			53	79.6
EFB_3/4			√		52	79.9
EFB_4/4				√	50	79.9

2.4 不同注意力模块对比实验

为了进一步检验DAB模块的有效性和合理性，本文选择了ECA-Net[18]在内的几种具有代表性的注意力机制与本文提出的DAB模块进行了对比实验，表6展示了不同注意力机制下的检测性能结果。

表6 不同注意力下的检测性能对比

Table 6 Comparison of detection performance under different attention

Method	FPS	mAP/%
SE	59	79.2
CBAM	57	79.7
ECA	60	79.6
Coordinate	52	79.8
DAB	52	**79.9**

从表6中可以看出，本文提出的DAB模块可以使得本模型的精度达到79.9%，对模型的检测精度的提升效果优于其他注意力模块，具有较大优势。

3 结论

在特征金字塔结构中如何进行有效的尺度变换和如何充分利用全局的上下文信息是提高检测性能的关键问题。本文针对此问题，提出一种引入残差学习与多尺度特征增强的目标检测器。首先在网络中引入基于残差学习的增强特征映射块，使得模型更加专注于对象区域而不是背景，并为有效特征图提高额外的语义信息；然后采用对上下文信息敏感的特征融合块，增大有效特征图的感受野，提高低分辨率下的检测性能；最后通过双重注意力块来抑制背景噪音，侧重于没有学习或者学习程度不足的小物体区域，进而提高模型的准确性和有效性。为了评估本文模型的性能，将RMFE-SSD与SSD和一些基于SSD改进的模型进行比较，并在PASCAL VOC和OAP两种数据集上进行测试。经过对实验结果对比分析得出，本文算法RMFE-SSD均有较大的精度优势，具有一定的应用价值和发展潜力。

参考文献：

[1] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks[C]// Advances in Neural Information Processing Systems 25, Lake Tahoe, Dec 3-6, 2012: 1097-1105.

[2] REN S, HE K, GIRSHICK R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[C]// Advances in Neural Information Processing Systems 28, Montreal, Dec 7-12, 2015: 91-99.

[3] ZHAO J, GUO W, ZHANG Z, et al. A coupled convolutional neural network for small and densely clustered ship detection in SAR images[J]. Science China Information Sciences, 2019, 62(4): 1-16.

[4] 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. Review of typical object detection algorithms for deep learning[J]. Computer Engineering and Applications, 2021, 57(8): 10-25.

[5] ZHANG K, ZHANG Z, LI Z, et al. Joint face detection and alignment using multitask cascaded convolutional networks [J]. IEEE Signal Processing Letters, 2016, 23(10): 1499-1503.

[6] WANG X Y, HAN T X, YAN S C. An HOG-LBP human detector with partial occlusion handling[C]//Proceedings of the IEEE 12th International Conference on Computer Vision, Kyoto, Sep 27-Oct 4, 2009. Washington: IEEE Computer Society, 2009: 32-39.

[7] LIN T Y, DOLLÁR P, GIRSHICK R B, et al. Feature pyramid networks for object detection[C]//Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, Jul 21-26, 2017. Washington: IEEE Computer Society, 2017: 936-944.

[8] KONG T, SUN F C, YAO A B, et al. RON: reverse connection with objectness prior networks for object detection[C]// Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, Jul 21-26, 2017. Washington: IEEE Computer Society, 2017: 5244-5252.

[9] LIU W, ANGUELOV D, ERHAN D, et al. SSD: single shot multibox detector[C]//LNCS 9905: Proceedings of the 14th European Conference on Computer Vision, Amsterdam, Oct 11-14, 2016. Cham: Springer, 2016: 21-37.

[10] LI Z, ZHOU F. FSSD: feature fusion single shot multibox

detector[J]. arXiv:1712.00960, 2017.
[11] SINGH B, DAVIS L S. An analysis of scale invariance in object detection SNIP[C]//Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake, Jul 18-22, 2018. Washington: IEEE Computer Society, 2018: 3578-3587.
[12] FU C Y, LIN W, RANGA A, et al. DSSD: deconvolutional single shot detector[C]//Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, Jul 21-26, 2017. Washington: IEEE Computer Society, 2017: 2881-2890.
[13] ZHOU P, NI B, GENG C, et al. Scale-transferrable object detection[C]//Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake, Jul 18-22, 2018. Washington: IEEE Computer Society, 2018: 528-537.
[14] 宋云博, 陈冬艳, 郝赟, 等. 基于级联卷积神经网络的高效目标检测方法[J]. 计算机工程与应用, 2021, 57(5): 139-145.
SONG Y B, CHEN D Y, HAO Y, et al. Efficient object detection method based on cascaded convolutional neural network[J]. Computer Engineering and Applications, 2021, 57(5): 139-145.
[15] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[J]. arXiv:1409.1556, 2014.
[16] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, Jun 27-30, 2016. Washington: IEEE Computer Society, 2016: 770-778.
[17] SZEGEDY C, IOFFE S, VANHOUCKE V, et al. Inception-v4, Inception-ResNet and the impact of residual connections on learning[C]//Proceedings of the 31st AAAI Conference on Artificial Intelligence, San Francisco, Feb 4-9, 2017. Menlo Park: AAAI, 2017: 4278-4284.
[18] 鞠默然, 罗江宁, 王仲博, 等. 融合注意力机制的多尺度目标检测算法[J]. 光学学报, 2020, 40(13): 132-140.
JU M R, LUO J N, WANG Z B, et al. Multi-scale target detection algorithm based on attention mechanism[J]. Acta Optica Sinica, 2020, 40(13): 132-140.
[19] WANG Q L, WU B B, ZHU P F, et al. ECA-Net: efficient channel attention for deep convolutional neural networks [C]//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, Jun 13-19, 2020. Piscataway: IEEE, 2020: 11531-11539.
[20] HU J, SHEN L, SUN G. Squeeze-and-excitation networks [C]//Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, Jun 18-22, 2018. Washington: IEEE Computer Society, 2018: 7132-7141.
[21] BELL S, ZITNICK C L, BALA K, et al. Inside-outside net: detecting objects in context with skip pooling and recurrent neural networks[C]//Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, Jun 27-30, 2016. Washington: IEEE Computer Society, 2016: 2874-2883.
[22] FASTER R. Towards real-time object detection with region proposal networks[C]//Advances in Neural Information Processing Systems 28, Montreal, Dec 7-12, 2015: 91-99.
[23] WOO S, PARK J, LEE J Y, et al. CBAM: convolutional block attention module[C]//LNCS 11211: Proceedings of the 15th European Conference on Computer Vision, Munich, Sep 8-14, 2018. Cham: Springer, 2018: 3-19.
[24] HOU Q B, ZHOU D Q, FENG J S. Coordinate attention for efficient mobile network design[C]//Proceedings of the 2021 IEEE Conference on Computer Vision and Pattern Recognition, Jun 19-25, 2021. Washington: IEEE Computer Society, 2021: 13713-13722.
[25] EVERINGHAM M, VAN G, WILLIAMS C, et al. The pascal visual object classes (VOC) challenge[J]. International Journal of Computer Vision, 2010, 88(2): 303-338.
[26] JEONG J, PARK H, KWAK N. Enhancement of SSD by concatenating feature maps for object detection[J].arXiv: 1705.09587, 2017.
[27] REDMON J, FARHADI A. YOLOV3: an incremental improvement[J]. arXiv:1804.02767, 2018.
[28] DAI J, LI Y, HE K, et al. R-FCN: object detection via region-based fully convolutional networks[C]//Advances in Neural Information Processing Systems 29, Barcelona, Dec 5-10, 2016: 379-387.

贾天豪(1996—),男,河北涿州人,硕士研究生,主要研究方向为深度学习、计算机视觉。
JIA Tianhao, born in 1996, M.S. candidate. His research interests include deep learning and computer vision.

彭力(1967—),男,河北唐山人,博士,教授,博士生导师,CAAI会员,CCF会员,主要研究方向为视觉物联网、行为识别、深度学习。
PENG Li, born in 1967, Ph.D., professor, Ph.D. supervisor, member of CAAI and CCF. His research interests include visual Internet of things, action recognition and deep learning.

戴菲菲(1988—),女,浙江临海人,硕士,工程师,主要研究方向为大数据、视觉物联网。
DAI Feifei, born in 1988, M.S., engineer. Her research interests include big data and visual Internet of things.