多尺度语义信息融合的目标检测

陈鸿坤 罗会兰*

(江西理工大学信息工程学院 赣州 341000)

摘 要：针对当前目标检测算法对小目标及密集目标检测效果差的问题，该文在融合多种特征和增强浅层特征表征能力的基础上提出了浅层特征增强网络(SEFN)，首先将特征提取网络VGG16中Conv4_3层和Conv5_3层提取的特征进行融合形成基础融合特征；然后将基础融合特征输入到小型的多尺度语义信息融合模块中，得到具有丰富上下文信息和空间细节信息的语义特征，同时把语义特征和基础融合特征经过特征重利用模块获得浅层增强特征；最后基于浅层增强特征进行一系列卷积获取多个不同尺度的特征，并输入各检测分支进行检测，利用非极大值抑制算法实现最终的检测结果。在PASCAL VOC2007和MS COCO2014数据集上进行测试，模型的平均精度均值分别为81.2%和33.7%，相对于经典的单极多盒检测器(SSD)算法，分别提高了2.7%和4.9%；此外，该文方法在检测小目标和密集目标场景上，检测精度和召回率都有显著提升。实验结果表明该文算法采用特征金字塔结构增强了浅层特征的语义信息，并利用特征重利用模块有效保留了浅层的细节信息用于检测，增强了模型对小目标和密集目标的检测效果。

关键词：目标检测；特征金字塔；特征融合；通道注意力；单极多盒检测器模型

中图分类号：TN911.73; TP391.4 **文献标识码**：A **文章编号**：1009-5896(2021)07-2087-09

DOI: 10.11999/JEIT200147

Multi-scale Semantic Information Fusion for Object Detection

CHEN Hongkun LUO Huilan

(*School of Information Engineering, Jiangxi University of Science and Technology, Ganzhou* 341000, *China*)

Abstract: Current object detection algorithms have poor detection results on small targets and dense targets. To address this challenge, a Shallow Enhanced Feature Network (SEFN) is proposed in this paper, which is based on the fusion of multiple features and enhanced shallow feature characterization capabilities. Firstly, the features extracted from the Conv4_3 layer and Conv5_3 layer are combined to form basic fusion features. Then the basic fusion features are inputted into a small multi-scale semantic information fusion module to obtain semantic features of rich contextual information and spatial detail information. The semantic features are fused into the basics features by the feature reuse module to obtain shallow enhanced features. Finally, a series of convolutions are performed based on the shallow enhanced features to obtain multiple features with different scales. Multiple detection branches are then constructed based on the features of different scales. The non-maximum suppression algorithm is used to achieve the final detection. The average accuracy of the proposed model is 81.2% and 33.7% on the PASCAL VOC2007 and MS COCO2014 datasets respectively, which is 2.7% and 4.9% higher than the classic Single Shot multibox Detector (SSD) algorithm. In addition, on detecting small targets in dense target scenes, the detection accuracy and recall rate of the proposed method are significantly improved. The experimental results show that the feature pyramid structure can enhance the semantic information of shallow features, and the feature reuse module can effectively retain shallow detail information for detection, so the proposed method can get better detection performance on small targets and dense targets.

收稿日期：2020-03-03；改回日期：2020-11-27；网络出版：2020-12-07

*通信作者： 罗会兰 luohuilan@sina.com

基金项目：国家自然科学基金(61862031, 61462035)；江西省教育厅科学技术研究项目(GJJ200859, GJJ200884)；江西省赣州市“科技创新人才计划”项目

Foundation Items: The National Natural Science Foundation of China (61862031, 61462035), The Science and Technology Research Project of Jiangxi Provincial Department of Education (GJJ200859, GJJ200884), Ganzhou City, Jiangxi Province “Technology Innovation Talent Program” Project

Key words: Object detection; Feature pyramid; Feature fusion; Channel attention; Single Shot multibox Detector (SSD) model

1 引言

目标检测是计算机视觉领域的热点研究课题，它的主要任务是对输入图像中的目标物体进行定位和分类。近年来，基于卷积神经网络的目标检测算法相对于传统算法取得了突破性的进展，已经得到广泛应用，但是由于检测图像中的小尺度目标仅具有少量的像素点，所包含的特征信息非常少，而且小尺度目标的特征经过多次下采样后容易丢失，使得特征提取网络很难提取到小尺度目标的特征，这就造成小尺度目标难以精确检测。而在密集目标检测场景中，不但存在大量小尺度目标，而且目标间有较严重的遮挡问题，使得密集目标的特征提取变得更加困难，所以小尺度目标和密集目标的检测仍然是目标检测领域中的难点和挑战。

图像中经常存在着大量尺度变化剧烈的目标物体，为了解决尺度变化问题，Liu等人[1]提出了单极多盒检测器(Single Shot multibox Detector, SSD)算法，其采用了多尺度特征检测思想，利用多个不同卷积层的不同尺度特征实现预测各种尺度的物体。但是，由于浅层特征缺乏语义信息，高层特征缺乏空间细节信息，SSD算法简单提取多尺度特征，没有考虑到不同尺度间的相关性，以至于不同层的语义信息[2]没有得到充分利用，所以其对小目标物体的检测准确度较低。为了更好地解决这个问题，文献[3]提出了特征金字塔网络(Feature Pyramid Network, FPN)，通过最近邻上采样操作额外构建了一个自顶向下的路径，并通过横向连接将高、低层中相同尺度的特征融合，有效减少了不同尺度特征间的语义信息差异，检测精度得到提高。在此基础上，基于SSD网络模型的改进研究DSSD[4]，利用反卷积操作构建FPN结构提取多尺度特征，获得了较高的检测精度，但是它的性能提升依赖于庞大的ResNet101[5]模型，导致检测效率大幅度下降。FSSD[6]算法和感受野模块网络(Receptive Field Block Network，RFB-Net)算法[7]在SSD算法的基础上进行改进，大幅度提高检测精度的同时保持了实时的检测速度。FSSD算法将浅层中3个不同尺度特征进行简单的拼接融合作为基准特征，并在基准特征的基础上进行一系列卷积下采样提取不同尺度特征，通过浅层特征间的信息融合，有效提高了检测器对小尺度目标的检测效果。RFB-Net则是利用3个不同扩张率的卷积层以Inception结构构建了RFB模块，增强网络的感受野，对大尺度物体的检测效果有很好的提升。上述方法都是在SSD算法的基础上，采用了不同的特征融合方式改进不同尺度特征的表征能力，从而提高模型对目标尺度变换的鲁棒性，虽然在一定程度上提升了模型对小物体的检测效果，但是没有着重关注浅层特征的作用，提取的浅层特征缺乏足够的语义信息，导致没有足够的能力来检测小目标，存在较为严重的小目标漏检和误检问题。

对于目标检测任务来说，卷积神经网络中的浅层特征保留了丰富的空间细节特征，对小尺度目标和密集目标的检测至关重要，但语义信息的缺乏限制了浅层特征的作用。为了提高小目标和密集目标检测性能，本文提出了一种新的融合多尺度语义信息以增强浅层特征的目标检测方法，旨在通过注意力机制强化有目标区域的浅层特征，从而获得更好的目标检测效果。本文的贡献主要体现在以下几个方面：

(1) 设计了多尺度语义信息融合模块对基础融合特征进行语义信息增强；

(2) 设计了特征重利用模块，通过语义特征指导基础融合特征进行通道注意力加权；

(3) 提出方法的检测精度在PASCAL VOC2007[8]数据集上相对于SSD512算法提高了2.7%，在MS COCO2014[9]数据集上相对于SSD512算法提高了4.9%。提出的浅层增强特征网络有效地提高了网络模型对小尺度目标和密集目标的检测效果，在MS COCO2014数据集上，对小目标检测精度相较于RFB512网络提高了1.6%，小目标召回率提高了1.8%。

2 本文方法

本文提出了结合多尺度语义信息和特征重利用来获取浅层增强特征的目标检测模型，简称为浅层特征增强网络(Shallow Enhanced Feature Network, SEFN)算法。具体框架如图1所示，主要设计了3个融合模块对浅层特征进行多尺度语义信息融合和特征重利用，从而获取浅层增强特征，并在此基础上提取多尺度特征进行检测，使模型在小目标检测上的性能获得显著提高。拼接融合模块是将多层浅层特征进行融合，获得具有更多详细空间细节信息的浅层特征，其通过将VGG16中Conv4_3层和Conv5_3层的特征进行拼接融合，得到基础融合特征。多尺度语义信息融合模块借鉴了语义分割任务中文献[10]的启发，采用特征金字塔结构使得基础融合特征引入了多尺度特征的语义信息，获得具有

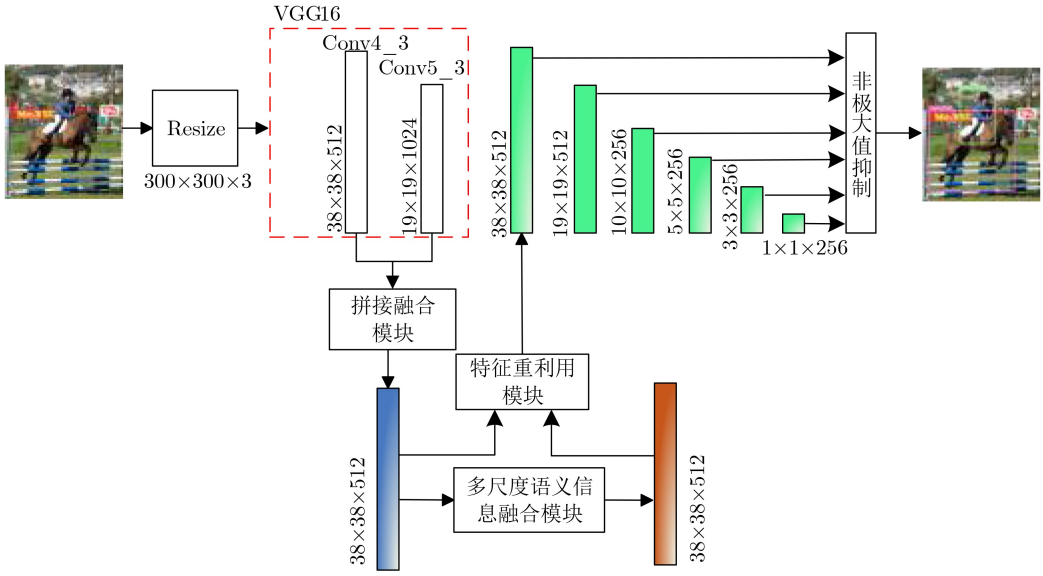

图 1 浅层特征增强网络

丰富上下文语义信息的语义特征，增强模型对小目标的分类能力。特征重利用融合模块则考虑到语义特征经过了一系列卷积和反卷积操作，会丢失部分细节纹理信息，所以重利用具有丰富空间细节信息的基础融合特征，并且依据语义特征进行通道注意力加权融合，更加关注包含目标的区域，有效提高了模型对小目标的检测精度和召回率。

2.1 浅层特征拼接融合模块

卷积神经网络的浅层特征包含丰富的纹理、边缘等细节空间信息，有利于对物体进行定位，但缺乏全局语义信息，不利于物体的分类。而高层特征具有丰富的语义信息，但丢失了许多空间细节信息，不利于小目标的检测。为了缓解这个问题，FSSD[6]算法通过简单拼接融合Conv4_3层、FC7层和Conv7_2层的特征，并在融合特征[11]的基础上进行卷积提取多尺度特征，因为融合特征建立了3个不同尺度特征间的联系，在一定程度上增加了语义信息和丰富了细节信息，所以在小目标检测效果上有一定的提升，但不同尺度间的联系过于单薄，增加的语义信息不足以很好地提高浅层特征的分类能力。本文在FSSD[6]算法思想的基础上，更加关注浅层特征的作用，首先设计了浅层特征的拼接融合模块，获得基础融合特征，如图2所示(示例中的输入图像尺寸为300×300)。浅层特征拼接融合模块的工作原理如式(1)所示

$$X_{\mathrm{f}}=\Phi_{\mathrm{f}}\left\{T_1\left(X_1\right),\ T_2(X_2)\right\} \tag{1}$$

其中，X_1, X_2表示选取的是VGG16网络中Conv4_3层和Conv5_3层卷积特征进行融合；T_1,T_2表示选取的特征在融合前进行的变换函数，将特征图变换到相同尺度以进行拼接融合，如图2所示，Conv4_3层提取的特征大小为38×38×512，而其变换函数是1×1的卷积层，目的是减少输入通道数，避免计算量过大，而Conv5_3层提取的特征大小为19×19×1024，其变换函数是具有双线性插值上采样的1×1卷积层，不仅降低输入通道数，而且统一了特征图的大小；下标f表示的是特征融合函数，本文采用的是拼接(Concatenation)操作；X_{f}表示获得的基础融合特征。

2.2 多尺度语义信息融合模块

通过简单拼接融合获取的基础融合特征，仅仅加强了空间细节特征，缺乏不同尺度间的语义信息，为了给浅层特征添加有效的多尺度语义信息，受语义分割研究领域中文献[10]提出的空间金字塔注意结构实现精确的像素级语义分割任务的启发，本文设计了一个类似特征金字塔结构的多尺度语义信息融合模块，结构如图3所示。

多尺度语义信息融合模块的输入是基础融合特征，由于输入特征的尺度较大，为了减少网络的计算负担，采用小型的1×1或3×3卷积核。该模块由3个分支组成，分别如图3中黄色、绿色和黑色的虚

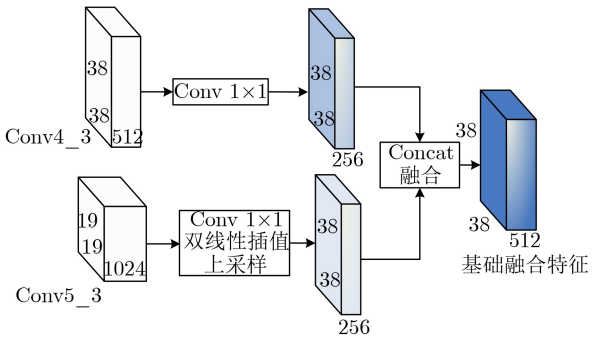

图 2 拼接融合模块

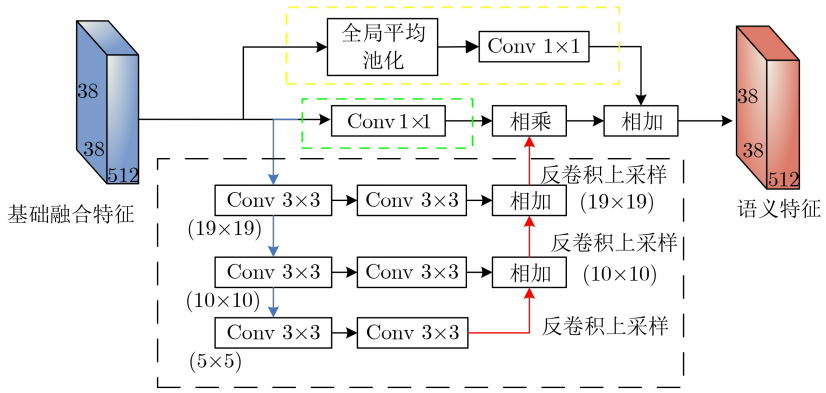

图 3 多尺度语义信息融合模块

线框，分别命名为全局语义分支、局部信息补充分支、多尺度语义分支。全局语义分支采用了一个全局平均池化层和一个1×1的卷积层，目的是提取基础融合特征的全局语义信息。局部信息补充分支只采用了一个1×1的卷积层，目的是为接下来的特征融合保留更多的空间细节信息。多尺度语义分支则构建了小型的特征金字塔结构，用于提取其他3个尺度特征的语义信息。在图3中黑色虚线框表示的多尺度语义分支中，蓝色的连线表示卷积下采样操作，红色的连线表示反卷积上采样操作，它们之间采用了卷积核大小为3×3，步长为2的卷积层进行连接，能够在不改变特征尺度的同时减少不同尺度间的语义相似性，随后采用对应元素相加操作逐步将3个不同尺度特征的语义信息进行融合。对局部信息补充分支的输出和多尺度语义分支的输出，采用对应元素相乘方式进行融合，使得特征不仅具有充分的空间细节信息，而且融合了多个尺度的语义信息。随后，为了更加关注图像物体的定位信息，对全局语义分支的全局语义信息，进行相加融合，最终得到了与输入特征相同尺度的语义特征。通过融合3个分支的不同信息，有效地对浅层特征嵌入了多个不同尺度的语义信息，加强对小目标检测的能力。多尺度语义信息融合模块的过程如式(2)所示

$$\left.\begin{aligned} X_{\mathrm{b1}} &= \mathrm{global}\left(X_{\mathrm{f}}\right) \\ X_{\mathrm{b2}} &= \mathrm{conv}\left(X_{\mathrm{f}}\right) \\ X_{\mathrm{b3}} &= \mathrm{fpa}(X_{\mathrm{f}}) \\ X_{\mathrm{s}} &= \mathrm{add}\left[X_{\mathrm{b1}}, \mathrm{multi}\left(X_{\mathrm{b2}}, X_{\mathrm{b3}}\right)\right] \end{aligned}\right\} \tag{2}$$

其中，X_{s}指的是模块最后获得的语义特征，X_{f}是拼接融合模块输出的基础融合特征，即本模块的输入特征，add ()表示对应像素相加的融合操作，multi ()表示对应像素相乘的融合操作，$X_{\mathrm{b1}}, X_{\mathrm{b2}}, X_{\mathrm{b3}}$分别表示全局语义分支、局部信息补充分支和多尺度语义分支获得的特征，而global (), conv (), fpa ()是这3个分支分别对基础融合特征进行的不同操作。

2.3 特征重利用模块

卷积神经网络中的高层特征具有丰富的类别语义信息，有利于指导低层特征进行通道注意力加权，使得具有较大空间分辨率的浅层更加关注存在目标的区域。借鉴文献[10]中全局注意力上采样模块的结构，本文设计了利用全局注意力信息的特征重利用模块，以进一步加强语义特征，结构如图4所示。首先对输入的语义特征进行全局平均池化操作，以获得各个特征通道的注意力权重，通过包含批处理归一化和ReLU操作的1×1卷积层后对基础融合特征相乘进行通道注意力加权，实现全局语义信息的增强。基础融合特征先经过了一个3×3卷积层处理，目的是保持与语义特征通道数一致，同时可以减少特征融合后的混叠效应，消除各级特征间特征分布的差异。最后将加权后的基础融合特征与语义特征进行对应元素相加融合，形成浅层增强特征。随后以浅层增强特征为基准，进行一系列卷积下采样操作，同时提取不同尺度的特征送入检测分支中进行检测(图1)。实验证明融合了不同尺度的语义信息，且具有丰富空间细节特征的浅层增强特征能够更加关注小目标的检测，不仅对小目标的检测精度有显著的提高，而且在密集目标检测中能取得好的效果。

2.4 损失函数

损失函数由定位损失函数和分类置信度损失函数两部分组成，表达式如式(3)所示：

$$L\left(x, c, l, g\right) = \frac{1}{N}\left(L_{\mathrm{conf}}(x, c) + \alpha L_{\mathrm{loc}}\left(x, l, g\right)\right) \tag{3}$$

其中，N是先验框的正样本数量，c为多类别目标的置信度预测值，l为先验框所对应边界框的位置预测值，即预测框与先验框的位置偏移量，g为真值框的位置参数；L_{conf}表示分类置信度损失函数，L_{loc}表示定位损失函数，α通过交叉验证设置为1。模型的定位损失函数是预测框偏移量l与真值框偏移量$\hat{g}$之间的$\mathrm{smooth}_{\mathrm{L1}}$损失，如式(4)所示，其中$\hat{g}$

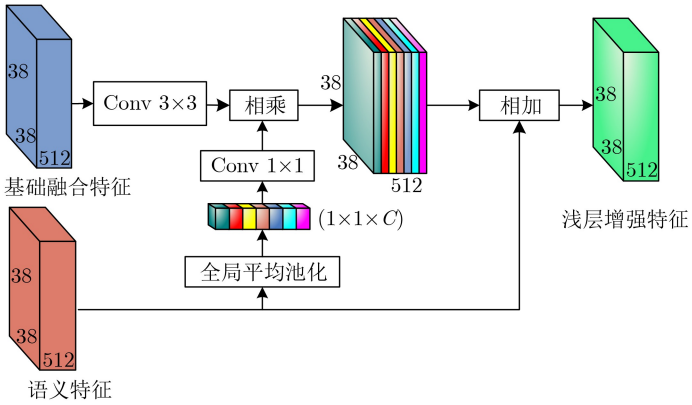

图 4 特征重利用模块

是指真值框g与先验框d之间的位置偏移量，$x_{ij}^k \in \{1,0\}$是用于将第i个先验框与类别k的第j个真值框进行匹配的指示符，若两框间的IOU大于或等于阈值0.5，则视为两框相匹配并且设置$x_{ij}^k=1$，反之$x_{ij}^k=0$。而(cx, cy)表示先验框的中心点坐标，w和h分别表示先验框的宽和高。Pos表示正样本集，GT表示真值标签集，num_class表示数据集中待检测类别总数。

$$L_{\mathrm{loc}}(x,l,g)=\sum_{i\in\mathrm{Pos}}\sum_{j\in\mathrm{GT}}\sum_{k=1}^{\mathrm{num\_class}}\sum_{m\in\{\mathrm{cx,cy},w,h\}} \cdot x_{ij}^k \mathrm{smooth}_{L1}\left(l_i^m-\hat{g}_j^m\right) \tag{4}$$

式(5)是为了获取真值框g与先验框d之间的位置偏移量$\hat{g}$，其中真值框位置参数为$g=(g^{\mathrm{cx}},g^{\mathrm{cy}},g^w,g^h)$，先验框位置参数为$d=(d^{\mathrm{cx}},d^{\mathrm{cy}},d^w,d^h)$。

$$\left.\begin{aligned}\hat{g}_j^{\mathrm{cx}}&=\left(g_j^{\mathrm{cx}}-d_i^{\mathrm{cx}}\right)/d_i^w\\ \hat{g}_j^{\mathrm{cy}}&=\left(g_j^{\mathrm{cy}}-d_i^{\mathrm{cy}}\right)/d_i^h\\ \hat{g}_j^w&=\ln\left(\frac{g_j^w}{d_i^w}\right)\\ \hat{g}_j^h&=\ln\left(\frac{g_j^h}{d_i^h}\right)\end{aligned}\right\} \tag{5}$$

式(6)为分类置信度损失函数，其是多类别置信度c上的softmax损失。

$$\left.\begin{aligned}L_{\mathrm{conf}}(x,c)&=-\sum_{i\in\mathrm{Pos}}\sum_{j\in\mathrm{GT}}\sum_{p=1}^{\mathrm{num\_class}}x_{ij}^p\ln(\hat{c}_i^p)-\sum_{i\in\mathrm{Neg}}\ln(\hat{c}_i^0)\\ \hat{c}_i^p&=\frac{\exp\left(c_i^p\right)}{\sum_p\exp(c_i^p)}\end{aligned}\right\} \tag{6}$$

3　实验结果

3.1　实验数据集

为了验证本文方法的有效性，在目标检测任务中常用的PASCAL VOC2007[8]和MS COCO2014[9]数据集上进行了实验分析。PASCAL VOC2007数据集包含20个常见类别，2501张训练图片、2510张验证图片和4952张测试图片。MS COCO2014数据集包含了80个类别，82783张训练图片、40504张验证图片和81434张测试图片，数据集中的目标大部分来自自然场景，目标间的尺度变化大，具有较多的小目标物体，评估标准也更加严格，要求算法具有更精确的定位能力。

3.2　实验设置

本文的模型是基于Pytorch1.0框架，Python版本为3.7，实验训练使用的是NVIDIA Tesla P100-PCIE-16 GB GPU, VGG16骨干网络的权重在ImageNet进行了预训练。采用随机梯度下降(SGD)算法来优化浅层特征增强网络的权重，其动量设置为0.9，衰减为0.0005，初始学习率为0.001。输入图像尺度设置为300×300时，训练PASCAL VOC数据集时设置每个GPU的batch size为32，训练MS COCO数据集时设置每个GPU的batch size为16。在输入图像尺度设置为512×512的情况下，训练PASCAL VOC数据集时设置每个GPU的batch size为16，训练MS COCO数据集时设置每个GPU的batch size为8。对于PSACAL VOC数据集，设置最大训练epoch为250代，在训练到150代和200代时，分别将学习率减少为原来的1/10；对于MS COCO数据集设置最大训练epoch为160代，在训练到90代和120代时，分别将学习率减少为原来的1/10。

采用了学习率热身算法来保证模型训练的稳定性，如式(7)所示。在训练刚开始时采用较小的学习率$1\times \mathrm{e}^{-6}$，并使用6个epoch来将学习率慢慢调回初始设置的学习率。式(7)中N_{iter}表示网络训练迭代的步数，一个batch size为一步，l_rate为网络初始的学习率，如0.001, epoch_size表示一代epoch有多少个batch size，即训练集中所有图片数量除以batch size的大小。

$$L\_\text{warm} = 1\times \text{e}^{-6} + \frac{N_{\text{iter}} \times \left(\text{l\_rate} - 1\times \text{e}^{-6}\right)}{\text{epoch\_size}\times 5} \quad (7)$$

3.3 实验结果分析

本小节比较了本文提出的浅层特征增强网络SEFN与近些年基于卷积神经网络的目标检测算法的性能，展现了本文提出的目标检测网络模型在小目标和密集目标检测方面的良好性能。

3.3.1 PASCAL VOC2007数据集上的检测性能对比

表1给出了近些年流行的目标检测算法与本文方法在PASCAL VOC2007数据集上的实验结果对比，这些方法都是联合使用PASCAL VOC2007和PASCAL VOC2012数据集中的训练集和验证集作为训练数据，在表1中“检测精度mAP(%), IOU=0.5”指的是在区分正负样本的区域交并比(Intersection Over Union, IOU)阈值选取在0.5的基础上，检测器获取的各类别平均精度的平均值，即平均精度均值(mean Average Precision, mAP)。“检测速度fps”表示的是检测器每秒能够处理的图片张数，其不仅与网络模型有关，而且与运行模型的硬件配置也相关。表1中算法简称后面的数字表示输入图像的尺度。

从表1的实验结果可以看出，本文提出的SEFN方法，在网络输入尺度为300×300时，获得了79.6%的检测精度，在网络输入尺度为512×512时，获得了81.2%的检测精度，能够在保证实时检测速度的同时，取得不错的检测效果。本文SEFN300方法与经典的2阶段目标检测算法R-FCN[12]取得了相当的检测精度，相较于经典的SSD300[1]算法提高了2.4%，与基于SSD算法的改进研究(如DSSD321[4], RSSD300[13], FSSD300[6])相比，分别提高了1%, 1.1%, 0.8%。而且相对一阶段流行的检测算法YOLOv1[14], YOLOv2[15]，在检测精度上有显著的提升，但相较于检测性能非常好的RFB300[7]算法，略低0.9%。本文的SEFN512方法的检测性能都明显高于表1中2阶段的目标检测算法，相较于经典的SSD512[1]算法提高了2.7%，较于RSSD512[13], kFSSD512[6]分别有0.4%，0.3%的提升，本文方法相较于DSSD512[4]略低0.3%，相较于RFB512[7]低1%，因为DSSD[4]采用了庞大的ResNet101[5]作为特征提取网络，其在512×512输入尺度下能够提取到表征能力更强的不同尺度特征进行检测，但在检测速度上牺牲很大，而RFB方法[7]则通过扩张卷积对多个不同尺度特征进行感受野增强，有效增强了大、中型尺度目标的检测能力，但本文方法在小目标和密集目标的检测效果具有明显的优势(如图5所示)。

表 1 在PASCAL VOC2007测试集本文方法与其他方法的结果对比

方法	骨干网络	输入尺度	GPU	fps(帧/s)	mAP(%),IOU=0.5
Faster RCNN[16]	VGG16	1000×600	Tian X	7.0	73.2
Faster RCNN[16]	ResNet-101	1000×600	K40	2.4	76.4
HyperNet[17]	VGG16	1000×600	Tian X	5.0	76.3
OHEM[18]	VGG16	1000×600	Tian X	7.0	74.6
ION[19]	VGG16	1000×600	Tian X	1.3	76.5
R-FCN[12]	ResNet-101	1000×600	K40	5.8	79.5
YOLOv1[14]	GoogleNet	448×448	Tian X	45.0	63.4
YOLOv2[15]	Darknet-19	352×352	Tian X	81.0	73.7
SSD300[1]	VGG16	300×300	Tian X	46.0	77.2
DSSD321[4]	ResNet-101	321×321	Tian X	9.5	78.6
RSSD300[13]	VGG16	300×300	Tian X	35.0	78.5
FSSD300[6]	VGG16	300×300	1080Ti	65.8	78.8
RFB300[7]	VGG16	300×300	1080Ti	83.0	80.5
本文SEFN300	VGG16	300×300	Tesla P100	55.0	79.6
YOLOv2[15]	Darknet-19	544×544	Tian X	40.0	78.6
SSD512[1]	VGG16	512×512	Tian X	19.0	78.5
DSSD513[4]	ResNet-101	513×513	Tian X	5.5	81.5
RSSD512[13]	VGG16	512×512	Tian X	16.6	80.8
FSSD512[6]	VGG16	512×512	1080Ti	35.7	80.9
RFB512[7]	VGG16	512×512	1080Ti	38.0	82.2
本文SEFN512	VGG16	512×512	Tesla P100	30.0	81.2

3.3.2 不同检测方法定性检测效果对比

如图5所示为本文方法与SSD[1]算法、RFB[7]算法在PASCAL VOC2007数据集上的图片检测效果对比。从图5的实验结果可以看出，SSD算法对小目标和密集目标的检测效果较差，经常出现严重的小物体误检和漏检的情况。RFB算法虽然在检测精度上高于本文的方法，但对小目标和密集目标的检测也存在较大不足，容易漏检小目标，并在密集目标的场景产生了目标误检问题，而本文采用多尺度语义信息增强浅层特征的方式，能够有效地加强对小目标和密集目标的检测性能，特别是模型在较大输入尺度的情况下(如512×512)。

3.3.3 MS COCO数据集上的检测性能对比

为了进一步体现本文方法在小目标和密集目标检测方面的优势，在MS COCO数据集的minival2014测试集上进行了本文方法与其他文献方法的实验结

(a) SSD300　(b) RFB300　(c) SEFN300

(d) SSD512　(e) RFB512　(f) SEFN512

图 5 不同算法在PASCAL VOC2007数据集上的检测结果

果对比，如表2所示，其中“IOU=0.5: 0.95”表示的是设定10个IOU阈值(0.5～0.95，以0.05为步长)，对每个IOU阈值求取算法的平均精度均值，再对所有IOU阈值对应的平均精度均值求平均。表中“S,M,L”分别表示小目标，中等目标，大目标。从表2的实验结果可以看出，本文提出的SEFN512方法在检测精度和召回率上相较于YOLOv2[15], SSD512[1], DSSD513[4], FSSD512[6]都有明显的提升，相比RFB512[7]，虽然总体检测精度低了0.7%，但本文的方法在小目标的检测精度上提高了1.6%，在中等目标上提高了1%，在小目标的召回率上提高了1.8%，有效证明了本文方法在小目标和密集目标检测上具有良好优势。本文方法主要关注浅层特征的作用，没有加强高层特征的表征能力，而高层特征的感受野相对较大，有利于大尺度目标的检测，RFB方法[7]更是增强了多个不同尺度特征的感受野，所以本文方法在大尺度目标检测上弱于RFB方法[7]，这是下一步改进工作的研究方向。

表 2 在MS COCO2014_minival测试集上本文方法与其他方法的结果对比

方法	骨干网络	检测精度mAP(%)			mAP(%)			召回率AR(%)		
		IOU=0.5:0.95	IOU=0.5	IOU=0.75	area: S	area: M	area: L	area: S	area: M	area: L
Faster R-CNN[16]	VGG16	24.2	45.3	23.5	7.7	26.4	37.1	–	–	–
R-FCN[12]	ResNet-101	29.2	51.5	–	10.3	32.4	43.3	–	–	–
YOLOv2[15]	Darknet-19	21.6	44.0	19.2	5.0	22.4	35.5	9.8	36.5	54.4
SSD512[1]	VGG16	28.8	48.5	30.3	10.9	31.8	43.5	16.5	46.6	60.8
DSSD513[4]	ResNet-101	33.2	53.3	35.2	13.0	35.4	51.5	21.8	49.1	66.4
FSSD512[6]	VGG16	31.8	52.8	33.5	14.2	35.1	45.0	22.3	49.9	62.0
RFB512[7]	VGG16	34.4	55.7	36.4	17.6	37.0	49.7	27.3	52.3	65.4
本文SEFN512	VGG16	33.7	54.7	35.6	19.2	38.0	47.3	29.1	52.5	63.2

4 结束语

为了更好地检测小目标和密集目标，本文提出了一种新的融合多尺度语义信息以增强浅层特征的目标检测方法。通过以SSD网络为基础框架，设计了3个不同的特征融合模块：拼接融合模块、多尺度语义信息融合模块和特征重利用模块。通过丰富多尺度语义信息和空间细节信息，使得浅层特征中包含目标的区域得到增强。最后在增强的特征上进行一系列的卷积提取多个不同尺度特征用以检测，有效地提高了对小目标和密集目标的检测性能。与主流的目标检测算法在PASCAL VOC2007数据集和MS COCO2014数据集上进行了实验结果对比，表明了本文方法能有效提高小目标和密集目标的检测效果。下一步工作将优化融合模块，从而更有效地利用浅层和高层特征之间的联系，使其获得更高的准确率。

参 考 文 献

[1] LIU Wei, ANGUELOV D, ERHAN D, *et al*. SSD: Single shot MultiBox detector[C]. The 14th European Conference on Computer Vision, Amsterdam, The Netherlands, 2016: 21–37.

[2] 罗会兰, 卢飞, 孔繁胜. 基于区域与深度残差网络的图像语义分割[J]. 电子与信息学报, 2019, 41(11): 2777–2786. doi: 10.11999/JEIT190056.
LUO Huilan, LU Fei, and KONG Fansheng. Image semantic segmentation based on region and deep residual network[J]. *Journal of Electronics & Information Technology*, 2019, 41(11): 2777–2786. doi: 10.11999/JEIT190056.

[3] LIN T Y, DOLLÁR P, GIRSHICK R, *et al*. Feature pyramid networks for object detection[C]. 2017 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, USA, 2017: 936–944.

[4] FU Chengyang, LIU Wei, RANGA A, *et al*. DSSD: Deconvolutional single shot detector[EB/OL]. http://arxiv.org/abs/1701.06659, 2017.

[5] HE Kaiming, ZHANG Xiangyu, REN Shaoqing, *et al*. Deep residual learning for image recognition[C]. 2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, 2016: 770–778.

[6] LI Zuoxin and ZHOU Fuqiang. FSSD: Feature fusion single shot multibox detector[EB/OL]. https://arxiv.org/abs/1712.00960, 2017.

[7] LIU Songtao, HUANG Di, and WANG Yunhong. Receptive field block net for accurate and fast object detection[C]. The 15th European Conference on Computer Vision, Munich, Germany, 2018: 404–419.

[8] 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. doi: 10.1007/s11263-009-0275-4.

[9] LIN T Y, MAIRE M, BELONGIE S, *et al*. Microsoft COCO: Common objects in context[C]. 13th European Conference on Computer Vision, Zurich, Switzerland, 2014: 740–755.

[10] LI Hanchao, XIONG Pengfei, AN Jie, *et al*. Pyramid attention network for semantic segmentation[C]. British Machine Vision Conference, Newcastle, UK, 2018.

[11] 罗会兰, 卢飞, 严源. 跨层融合与多模型投票的动作识别[J]. 电子与信息学报, 2019, 41(3): 649–655. doi: 10.11999/JEIT180373.

LUO Huilan, LU Fei, and YAN Yuan. Action recognition based on multi-model voting with cross layer fusion[J]. *Journal of Electronics & Information Technology*, 2019, 41(3): 649–655. doi: 10.11999/JEIT180373.

[12] DAI Jifeng, LI Yi, HE Kaiming, *et al*. R-FCN: Object detection via region-based fully convolutional networks[C]. The 30th International Conference on Neural Information Processing Systems, Barcelona, SPAIN, 2016: 379–387.

[13] JEONG J, PARK H, and KWAK N. Enhancement of SSD by concatenating feature maps for object detection[C]. British Machine Vision Conference, London, UK, 2017.

[14] REDMON J, DIVVALA S, GIRSHICK R, *et al*. You only look once: Unified, real-time object detection[C]. 2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, 2016: 779–788.

[15] REDMON J and FARHADI A. YOLO9000: Better, faster, stronger[C]. 2017 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, USA, 2017: 6517–6525.

[16] REN Shaoqing, HE Kaiming, 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*, 2017, 39(6): 1137–1149. doi: 10.1109/TPAMI.2016.2577031.

[17] KONG Tao, YAO Anbang, CHEN Yurong, *et al*. HyperNet: Towards accurate region proposal generation and joint object detection[C]. 2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, 2016: 845–853.

[18] SHRIVASTAVA A, GUPTA A, and GIRSHICK R. Training region-based object detectors with online hard example mining[C]. 2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, 2016.

[19] BELL S, ZITNICK C L, BALA K, *et al*. Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks[C]. 2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, 2016, 2874–2883.

陈鸿坤：男，1995年生，硕士，研究方向为目标检测.

罗会兰：女，1974年生，教授，博士后，主要研究方向为机器学习、模式识别.

责任编辑：马秀强