激光与光电子学进展 先进成像

基于单次双向特征金字塔网络的目标检测模型

张云川，姜麟*，林莉
昆明理工大学理学院，云南 昆明 650500

摘要 目标检测是计算机视觉领域的重点研究方向，SSD模型虽然在检测精度与速度上均取得较好的效果，但其利用低语义信息的浅层特征训练小目标，容易出现小目标漏检和误检现象。对此，提出了一种基于单次双向特征金字塔网络的改进SSD目标检测模型(OBSSD)。首先基于分层融合的思想构建双向特征融合模块来解决浅层特征利用不足的问题；其次引入融合权重来更加有效地融合不同层级的特征，改善浅层特征语义信息低的问题；最后在分类和回归预测前加入基于残差模块的检测单元，解决因分类网络偏向平移不变性导致的目标定位不准确问题。在PASCAL VOC2007测试集上的实验结果表明：所提模型的平均精确度(mAP)为80.8%，较SSD模型提高6.5个百分点，且检测速度满足实时检测的需求。

关键词 机器视觉；目标检测模型；特征融合；语义信息；实时性
中图分类号 TP391 **文献标志码** A **DOI**：10.3788/LOP220555

Target Detection Model Based on Once Bidirectional Feature Pyramid Network

Zhang Yunchuan, Jiang Lin*, Lin Li
Faculty of Science, Kunming University of Science and Technology, Kunming 650500, *Yunnan, China*

Abstract Target detection is an important research direction in the field of computer vision. Although the single-shot detector (SSD) model achieves good results in terms of detection accuracy and speed, its use of shallow features with low semantic information for training small targets is prone to target misses and false detections. In this paper, an improved SSD target detection model based on a once bidirectional feature pyramid network (OBSSD) is proposed. First, a bidirectional feature fusion module is constructed based on the principle of hierarchical fusion to solve the problem of under-utilization of shallow features. Second, fusion weights are introduced to fuse features at different levels more effectively and mitigate the problem of low semantic information of shallow features. Finally, a detection unit based on the residual module is added before classification and regression prediction to address the problem of inaccurate target localization caused by the biased translational invariance of the classification network. The experimental results on the PASCAL VOC2007 test set show that the mean average precision (mAP) of the proposed model is 80.8%, which is 6.5 percentage points higher than that of the SSD model, and the detection speed meets the demand for real-time detection.

Key words machine vision; target detection model; feature fusion; semantic information; real-time

1 引 言

目标检测任务一直是计算机视觉领域的焦点问题[1]，而小目标检测是目标检测任务中的难点。小目标检测的准确度低是由于神经网络不能够有效学习图像中所占比例较小且所含信息不足的目标特征。传统目标检测模型因存在检测效率低、精确性差和泛化能力弱等问题而逐渐被基于深度学习的高性能目标检测模型所代替。现阶段，目标检测模型可分为基于分类的两阶段模型和基于回归的单阶段模型[2-3]。两阶段Faster R-CNN[4]目标检测模型通过搜索算法生成候选框集，然后利用卷积神经网络提取特征并对特征进行有效分类和定位，但实时性能较差。孙跃军等[5]改进Mask R-CNN[6]的基准网络，在新网络Mask R-CNN-Ⅱ中应用迁移学习算法，有效提升了目标的定位与分类精准性。Redmon等[7]提出了一种检测速度更快的单阶段YOLO目标检测模型，但YOLO对近距离物体的检测效果不好，且网络

收稿日期：2022-01-17；**修回日期**：2022-02-25；**录用日期**：2022-03-14；**网络首发日期**：2022-03-26
基金项目：国家自然科学基金项目(11761042)、云南省教育厅基金(KKJB201707008)
通信作者：*tojianglin@126.com

泛化能力弱。张官荣等[8]以通道的重要性为依据对YOLOv3-CS[9]进行剪枝，在损失少许检测精确度的情况下，大幅度压缩模型的大小，有效提升了检测速度。

Liu等[10]提出的SSD模型实现了检测精度和速度上的相对平衡，但该模型使用较低层级的锚框训练小目标，浅层特征的低语义信息导致模型对小目标的检测效果偏差。Fu等[11]替换SSD的主干网络，添加反卷积层形成“沙漏”结构来提升小目标检测性能，但主干网络参数量过大，导致检测速度急剧降低。Jeong等[12]利用分类网络加强不同层特征图的联系来减少重复框出现的概率，增加特征金字塔中特征图的数量来检测更多的小目标。陈幻杰等[13]对浅层特征层采用区域放大提取法来提高模型对小目标的检测性能。周永福等[14]通过增加深度通道网络形成双通道SSD模型来克服单一图像的检测弊端。李青援等[15]通过建立全局像素点之间的长距离关系和各通道之间的重要性关系来有效提升模型检测精度。郭瑞鸿等[16]修改SSD中的基础网络，对基础网络后三层卷积进行特征融合，改进拓展层构建新的网络结构为小目标检测提供充分的上下文语义信息。Yin等[17]提出了一种基于特征融合和空洞卷积的单步多框检测器，多层特征融合模块和多分支残差扩张卷积模块有效提升了小目标检测精确度。

本文从SSD模型存在的小目标检测精度低和相关改进模型实时性差等问题入手，提出了一种基于单次双向特征金字塔(Once Bi-FP)网络的改进SSD目标检测模型(OBSSD)，以便更好地权衡检测精度与检测速度。首先，将单次双向特征金字塔模块置于主干网络与预测网络之间，充分利用浅层特征中有利于准确预测物体形状、位置的细节信息和深层特征中的高语义信息，加强浅层特征提取；其次，引入特征融合权重对高层特征和低层特征进行更加有效融合，丰富低层特征的语义信息；最后，在分类和回归预测前引入基于残差模块的检测单元，避免使用偏向平移不变性的主干网络直接对特征进行分类和识别，提高目标定位准确性。

2　改进SSD模型

2.1　SSD模型

SSD模型将YOLO模型中的回归思想和Faster R-CNN中的锚框思想相结合，利用卷积神经网络提取特征，均匀地在图片不同位置进行不同尺度和长宽比的密集抽样，且物体分类和预测框回归同时进行，使其保证检测精度的同时兼顾检测速度。但是SSD模型用较低层级的特征来训练小目标，小目标即物体宽高是原图宽高$\frac{1}{10}$以下的目标。浅层特征非线性程度不够，所含的语义信息就不能让网络学习到足够的有用特征，导致模型对小目标识别效果差。SSD模型以改进的VGG16[18]网络和额外添加的8层卷积层作为主干特征提取网络，将VGG16的fc6和fc7转化为卷积层，去除所有的Dropout层和fc8层。SSD模型框架如图1所示。

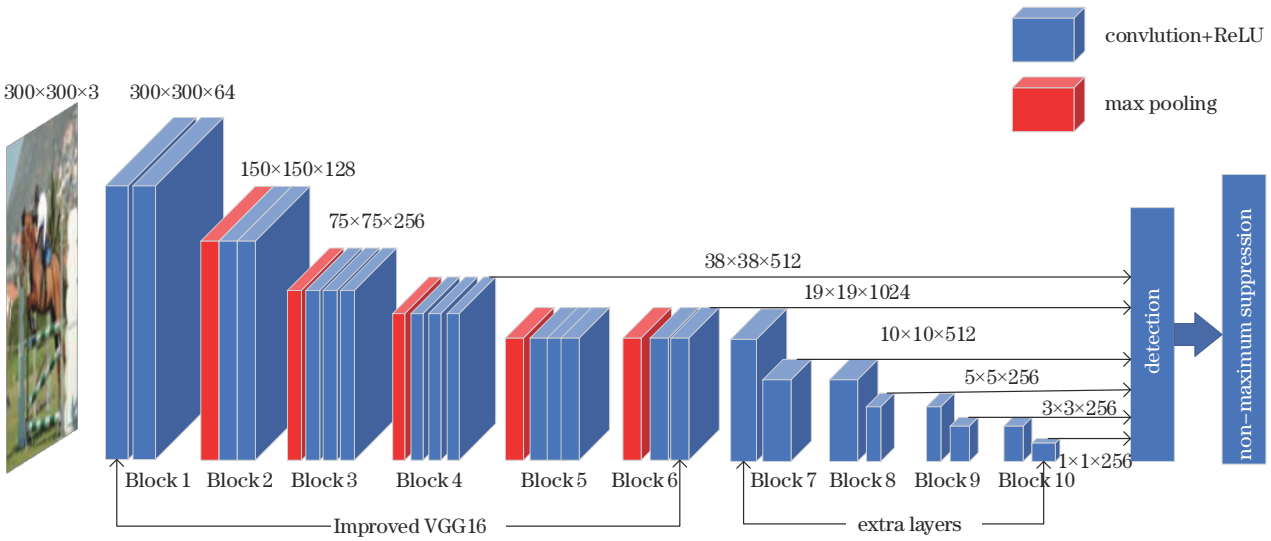

图1　SSD模型框架

Fig. 1　SSD model framework

图1中，SSD模型将主干网络提取到的特征直接用于物体分类和定位检测，通过最大池化和卷积操作逐步降低特征图尺寸，从网络不同层提取不同尺度的特征，然后对特征分别进行检测，完成物体的分类和位置坐标的回归。SSD模型主干网络结构的相关参数如表1所示，其中Conv表示卷积，Act表示激活函数，MaxPooling表示最大池化，k表示卷积核或池化核的尺寸大小，s表示步长，p表示填充像素大小，d表示膨

表 1 SSD 主干网络结构

Table 1 SSD backbone network structure

Block	Layer	Operation	Specific operational detail	Output feature size
Block 1	Conv1_1	Conv,Act	$k=3,p=1$;ReLU	300×300×64
	Conv1_2	Conv,Act	$k=3,p=1$;ReLU	300×300×64
Block 2	Pooling1	MaxPooling	$k=2,s=2$	150×150×64
	Conv2_1	Conv,Act	$k=3,p=1$;ReLU	150×150×128
	Conv2_2	Conv,Act	$k=3,p=1$;ReLU	150×150×128
Block 3	Pooling2	MaxPooling	$k=2,s=2$	75×75×128
	Conv3_1	Conv,Act	$k=3,p=1$;ReLU	75×75×256
	Conv3_2	Conv,Act	$k=3,p=1$;ReLU	75×75×256
	Conv3_3	Conv,Act	$k=3,p=1$;ReLU	75×75×256
Block 4	Pooling3	MaxPooling	$k=2,s=2$	38×38×256
	Conv4_1	Conv,Act	$k=3,p=1$;ReLU	38×38×512
	Conv4_2	Conv,Act	$k=3,p=1$;ReLU	38×38×512
	Conv4_3	Conv,Act	$k=3,p=1$;ReLU	38×38×512
Block 5	Pooling4	MaxPooling	$k=2,s=2$	19×19×512
	Conv5_1	Conv,Act	$k=3,p=1$;ReLU	19×19×512
	Conv5_2	Conv,Act	$k=3,p=1$;ReLU	19×19×512
	Conv5_3	Conv,Act	$k=3,p=1$;ReLU	19×19×512
Block 6	Pooling5	MaxPooling	$k=2,s=1,p=1$	19×19×512
	Conv6	Conv,Act	$k=3,p=6,d=6$;ReLU	19×19×1024
	Conv7	Conv,Act	$k=1$;ReLU	19×19×1024
Block 7	Conv8_1	Conv,Act	$k=1$;ReLU	19×19×256
	Conv8_2	Conv,Act	$k=3,s=2,p=1$;ReLU	10×10×512
Block 8	Conv9_1	Conv,Act	$k=1$;ReLU	10×10×128
	Conv9_2	Conv,Act	$k=3,s=2,p=1$;ReLU	5×5×256
Block 9	Conv10_1	Conv,Act	$k=1$;ReLU	5×5×128
	Conv10_2	Conv,Act	$k=3,p=1$;ReLU	3×3×256
Block 10	Conv11_1	Conv,Act	$k=1$;ReLU	3×3×128
	Conv11_2	Conv,Act	$k=3,p=1$;ReLU	1×1×256

胀系数。

从表 1 可以看出，主干网络由卷积、激活函数和最大池化构建而成，激活函数均采用 ReLU[19]。其原因是对于线性函数而言，ReLU 的表达能力更强，尤其体现在深度网络中；对于非线性函数而言，ReLU 由于非负区间的梯度为常数，因此不存在梯度消失问题，使得模型的收敛速度维持在一个稳定状态，且 ReLU 函数表达式简单，可加快网络收敛速度。Conv6 采用空洞卷积[20]，其目的在于增大网络感受野，利用较大的感受野和低层特征来提高模型的定位能力，增强检测目标的位置回归准确性。

通过主干网络，将提取到的 6 个不同尺度的特征层作为有效特征层进行下一步处理，每个有效特征层将整张图像分成与其长宽对应的网格，然后从每个网格的中心点建立多个先验框。各有效特征层上的网格所含先验框数量的对应关系如表 2 所示。

表 2 有效特征层上单个网格先验框数量

Table 2 Number of prior frames of a single grid on effective feature layer

Efficient feature layer	Size	Number of prior frames per grid
Conv4_3	38×38	4
Conv7	19×19	6
Conv8_2	10×10	6
Conv9_2	5×5	6
Conv10_2	3×3	4
Conv11_2	1×1	4

以 Conv4_3 这一有效特征层为例，将整张图像分成 38×38 即 1444 个网格，然后从每个网格的中心点建立 4 个先验框，所以该有效特征层共有 5776 个先验框。同理，其余有效特征层的先验框数量可通过类似

计算得到，整个模型共产生8732个先验框。

SSD模型存在需要预先设置的参数，如先验框的数量、大小及长宽比，导致模型调试过程依赖于经验，所提模型也同样存在这一问题。为了控制变量，所提模型在如上参数的设置上与SSD模型保持一致。

2.2　OBSSD模型

SSD模型通过下采样特征图以实现多尺度检测，并将提取到的6种尺度的特征层作为待检测层，用较大的特征图检测相对较小的目标，较小的特征图检测较大的目标，最后以卷积的方式实现物体的分类和位置回归。但SSD模型没有融合具有丰富细节信息的浅层特征和高语义信息的深层特征，导致对小目标检测精度较低。针对SSD模型存在的问题，所提模型利用单次双向特征金字塔[21]模块加强浅层特征提取，引入融合权重来更加有效地融合浅层特征和深层特征，增强浅层特征的语义信息。图2为所提模型框架图。

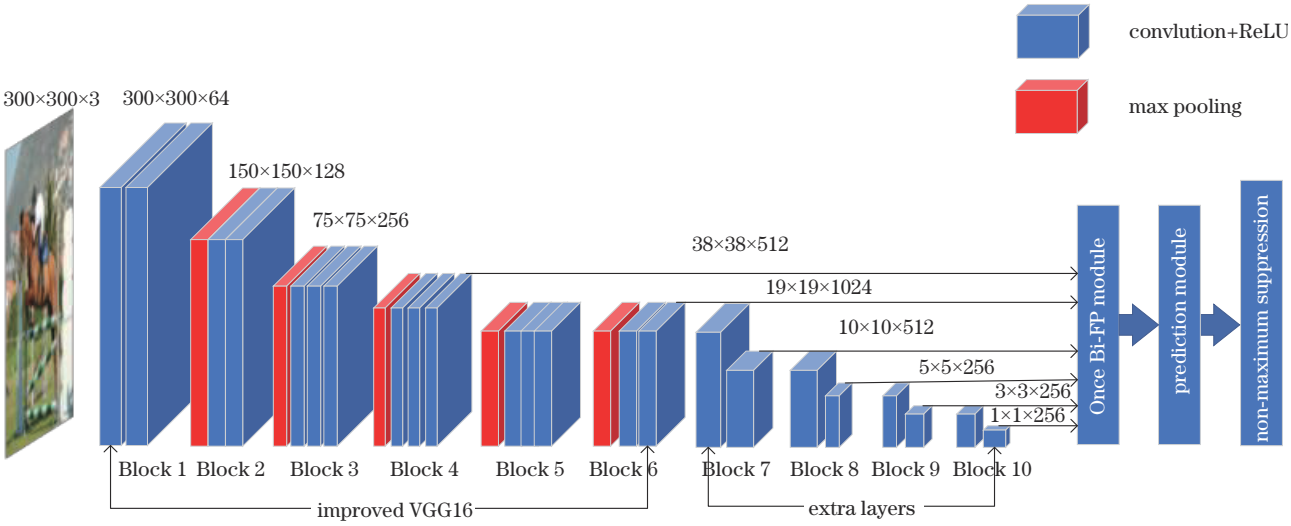

图2　所提模型框架

Fig. 2　Proposed model framework

图片先经主干网络提取特征，随后将6种不同尺度的特征层传入单次双向特征金字塔模块，之后再将分层融合得到的特征传入基于残差的预测模块提取更深维度的特征用于目标分类和回归，提高模型检测精确度。

2.2.1　单次双向特征金字塔模块

为改进SSD模型对浅层特征利用不足的问题，将单次双向特征金字塔模块置于主干网络和预测网络之间，加强浅层特征提取，引入特征融合权重分层融合浅层特征与深层特征，学习不同特征的重要性。通过多尺度融合低层特征和高层特征，有效改善浅层特征语义信息不足的问题。图3为单次双向特征金字塔模块结构图。

单次双向特征金字塔模块采用从上到下、自底向顶分层融合的方式融合浅层特征和深层特征，充分利用浅层特征中利于目标回归的细节信息和深层特征中利于目标分类的高语义信息，加强浅层特征提取，增加浅层特征的语义信息，提高小目标检测精度。由于数据集中包含小目标的图片较少且物体本身占原图像比例小，小目标对损失函数的贡献少，模型在训练的时候偏向中等和大目标，故浅层的细节信息更有助于小目标检测。以往的特征金字塔网络(FPN)[22]和路径增强网络(PA-Net)[23]在融合特征层时，均对特征进行简单相加，未考虑输入特征图分辨率的不同对融合后输出特征的贡献度不一样。因此，所提单次双向特征金字塔模块在进行特征融合时引入可学习的权重，目的是确定在融合过程中网络更关注哪一个特征。

2.2.1.1　从上到下的特征融合过程

先对P_{i+1_im}上采样到与P_{i_in}相同的尺寸大小，再将P_{i+1_im}和P_{i_in}进行加权特征融合并利用Swish函数对融合后的特征进行激活，随后用深度可分离卷积[24]进行特征提取，减少参数数量和运算成本，并在卷积后添加批归一化(BN)和Swish[25]激活函数，最后输出P_{i_im}。从上到下的融合模块如图4所示。

$P_{i_\mathrm{im}}(i=1,\cdots,4)$表示从上到下路径中第$i$级的中间特征，$P_{i_\mathrm{in}}(i=1,\cdots,4)$表示有效特征序列的第$i$级输入特征。$P_{i_\mathrm{im}}$的表达式为

$$P_{i\_\mathrm{im}}=\mathrm{Swish}\left\{\mathrm{BN}\left\{\mathrm{Conv}\left\{\mathrm{Swish}\left[\frac{w_1\times P_{i\_\mathrm{in}}+w_2\times \mathrm{upsample}(P_{i+1\_\mathrm{im}})}{w_1+w_2+\varepsilon}\right]\right\}\right\}\right\},\tag{1}$$

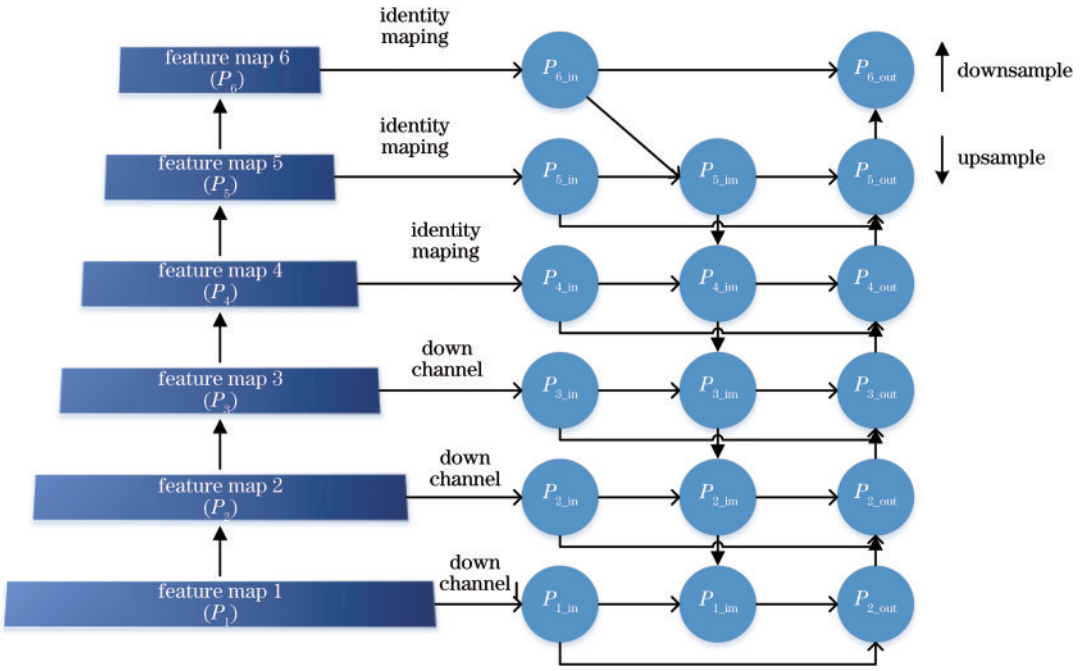

图 3　单次双向特征金字塔模块

Fig. 3　Once Bi-FP module

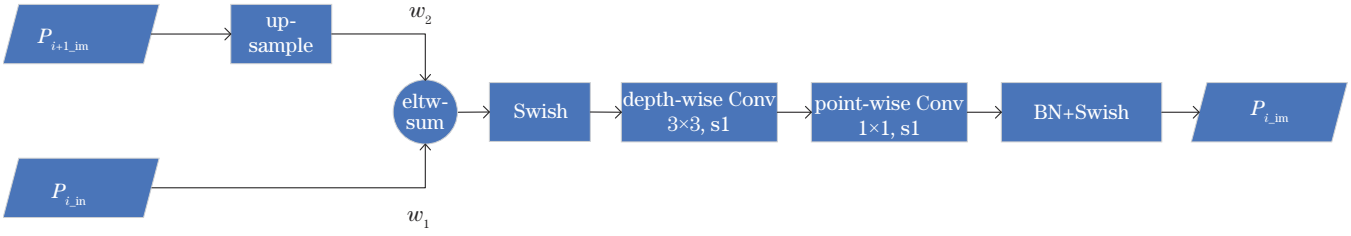

图 4　从上到下特征融合模块

Fig. 4　Top to bottom feature fusion module

式中：特征融合权重 $w_k \geqslant 0(k=1,2)$；ε 取 0.001 来保证特征融合过程的数值稳定性。P_{5_im} 的表达式为

$$P_{5\_im}=\text{Swish}\left\{\text{BN}\left\{\text{Conv}\left\{\text{Swish}\left[\frac{w_1\times P_{5\_in}+w_2\times \text{upsample}(P_{6\_in})}{w_1+w_2+\varepsilon}\right]\right\}\right\}\right\}。\tag{2}$$

利用 Swish 函数对加权融合后的特征进行激活，增加特征的非线性特性；再通过深度可分离卷积提取特征的丰富属性；然后将特征批归一化，加快网络收敛；最后再对特征进行激活，避免梯度消失。

2.2.1.2　自底向顶的特征融合过程

先对 P_{i-1_out} 下采样到与 P_{i_im} 相同的尺寸大小，再将 P_{i_in}、P_{i_im} 和 P_{i-1_out} 进行加权特征融合并利用 Swish 函数对融合后的特征进行激活，随后用深度可分离卷积进行特征提取，同样在卷积后添加批归一化和 Swish 激活函数，最后得到 P_{i_out}。自底向顶的融合模块如图 5 所示。

$P_{i\_out}(i=2,\cdots,5)$ 表示自底向顶路径中第 i 级的输出特征，P_{i_out} 的表达式为

$$P_{i\_out}=\text{Swish}\left\{\text{BN}\left\{\text{Conv}\left[\text{Swish}\frac{w_1'\times P_{i\_in}+w_2'\times P_{i\_im}+w_3'\times \text{downsample}(P_{i-1\_out})}{w_1'+w_2'+w_3'+\varepsilon}\right]\right\}\right\},\tag{3}$$

式中：特征融合权重 $w_j' \geqslant 0(j=1,2,3)$；同样 ε 取 0.001 来保证特征融合过程的数值稳定性。P_{1_out} 和 P_{6_out} 的表达式分别为

$$P_{1\_out}=\text{Swish}\left\{\text{BN}\left\{\text{Conv}\left[\text{Swish}\left(\frac{w_1'\times P_{1\_in}+w_2'\times P_{1\_im}}{w_1'+w_2'+\varepsilon}\right]\right\}\right\}\right\},\tag{4}$$

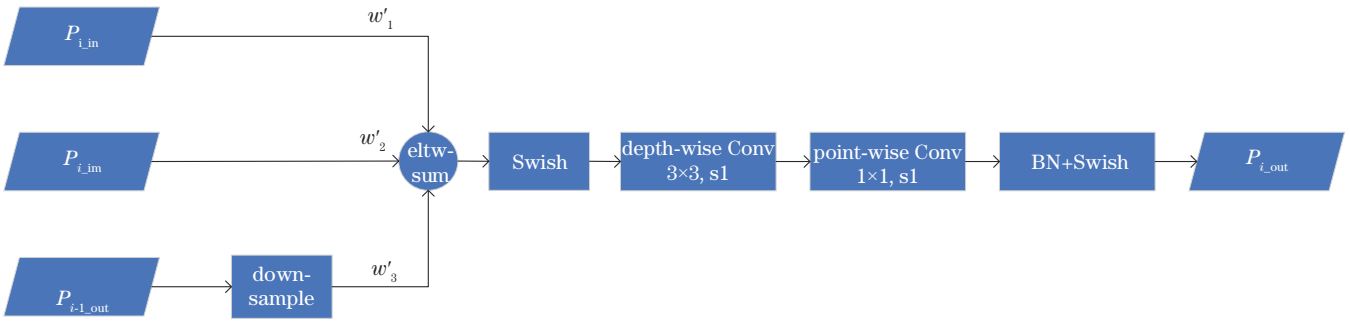

图 5　自底向顶特征融合模块

Fig. 5　Bottom-to-top feature fusion module

$$P_{6\_out}=\mathrm{Swish}\left\{\mathrm{BN}\left\{\mathrm{Conv}\left\{\mathrm{Swish}\left[\frac{w'_1\times P_{6\_in}+w'_2\times \mathrm{downsample}(P_{5\_out})}{w'_1+w'_2+\varepsilon}\right]\right\}\right\}\right\}。\tag{5}$$

2.2.2　预测模块

目标检测模型的主干网络常由分类网络构成，而分类网络偏向平移不变性，忽略了位置信息。因此，直接使用主干网络提取到的特征进行预测会影响目标定位精准性。对此，所提模型在分类和回归预测前加入残差模块的检测单元[26]，避免直接使用主干网络提取到的特征检测目标，同时提取更深维度的特征用于目标分类和回归。图 6 为预测模块结构图。

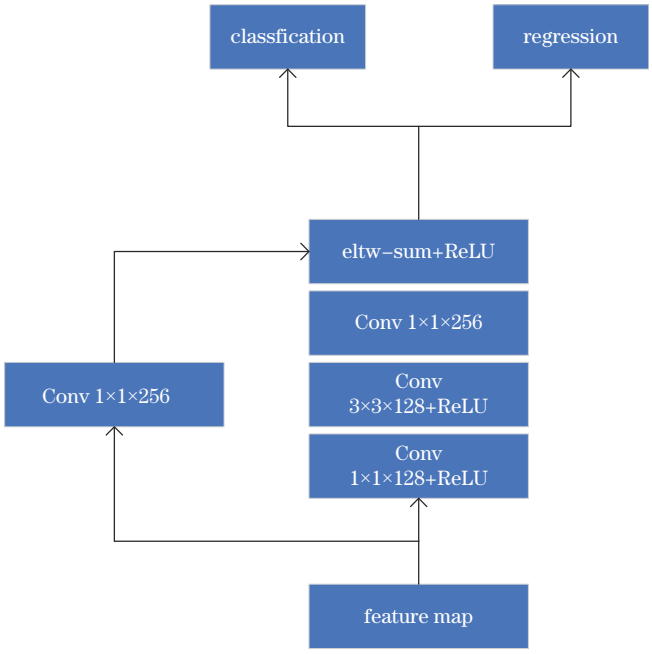

图 6　预测模块

Fig. 6　Prediction module

将单次双向特征金字塔模块的输出特征作为预测模块的输入特征。预测模块的主分支上，特征先经尺寸大小为 1×1 的卷积核降低通道数，然后用 3×3 的卷积核进行特征整合，接着再由 1×1 的卷积核恢复通道数；残差分支由一个 1×1 的卷积核构成，目的是统一残差分支输出特征和主分支输出特征的维度；最后将主分支和残差分支的特征按逐元素相加的方式融合，随后将融合后的特征进行 ReLU 激活，增加特征的非线性特性。卷积后面加非线性激活函数可降低预测模块产生的梯度值对主干特征提取网络的影响[26]，增强网络层间的非线性关系，避免在网络训练后期出现梯度消失，在一定程度上还可增加网络的稀疏性，防止网络过拟合。

3　实验结果与分析

3.1　实验准备

3.1.1　实验环境

实验编程语言为 Python 3.7，深度学习框架为 PyTorch 1.5，操作系统为 Ubuntu 18.04，独立显卡为 NVIDIA GeForce GTX 1660 SUPER，CUDA 版本为 10.1。

3.1.2　数据集

模型在公开数据集 PASCAL VOC2007 和 PASCAL VOC2012[27] 上进行训练与测试，该数据集包括人、交通工具、动物和室内物体等 4 个大类，共计 20 个小类。PASCAL VOC2007 共有 9963 张图片（包含 24640 个目标），其中训练集有 5011 张图片，测试集有 4952 张图片；PASCAL VOC2012 共有 23080 张图片（包含 54900 个目标），其中训练集有 11540 张图片，测试集有 11540 张图片。在 PASCAL VOC2007 训练集上联合 PASCAL VOC2012 训练集（包含 16551 张图片，40058 个目标）进行训练，然后使用 PASCAL VOC2007 测试集（包含 4952 张图片，12032 个目标）进行测试。

3.1.3　训练策略

实验训练过程分为两个阶段，如表 3 所示，其中 Lr 是学习率的缩写。

第 1 阶段：以在 ImageNet[28] 数据集上训练好的 VGG16 作为预训练模型对 SSD 模型进行训练，先对输入图片进行数据预处理，采用图像翻转、长宽扭曲和

表 3 训练策略

Table 3 Training strategies

Stage	Optimizer	Batch_size	Freeze_train	Initial_Lr	Lr_scheduler	Epoch
1	Adam	32	True	0.0005	ReduceLROnPlateau	50
	Adam	16	False	0.0001	ReduceLROnPlateau	150
2	SGD-M	32	True	0.001	MultiStepLR	50
	SGD-M	16	False	0.001	MultiStepLR	50

颜色扰动等数据增强操作，利用 Adam 优化器，根据损失值的变化情况自适应调整学习率，容忍轮次为 5，调整因子为 0.5。首先冻结预训练模型，初始学习率设为 0.0005，批次大小设为 32，训练 50 个轮次；然后解冻训练，以 0.0001 为初始学习率，批次大小设为 16，训练 150 个轮次。共训练 200 个轮次得到第 1 阶段的模型。

第 2 阶段：以第 1 阶段训练得到的模型作为本阶段的预训练模型，首先冻结预训练模型，对所增加的网络结构层进行训练，批次大小设为 32，利用 SGD-M 优化器，动量为 0.9，按设定的轮次间隔调整学习率。初始学习率设为 0.001，训练至 40 轮次时学习率下降为原来的 1/10，共训练 50 个轮次；然后微调整个网络，批次大小设为 16，同样利用 SGD-M 优化器，动量为 0.9，按设定的轮次间隔调整学习率。初始学习率设为 0.001，训练至 30 轮次、40 轮次时学习率依次下降为原来的 1/10，训练 50 个轮次。此阶段共训练 100 个轮次得到最终模型。

3.1.4 评估标准

实验选取目标检测中评估模型的通用标准：平均精确度（mAP）和每秒帧率（FPS）。mAP 是所有类平均准确率（AP）的平均值，用来评估模型的检测精准性，FPS 是每秒处理图像的帧数，用来衡量模型的检测速度。

3.2 实验结果

对比了多种主流模型在 PASCAL VOC2007 测试集上的 mAP 和 FPS，具体结果如表 4 所示。

表 4 PASCAL VOC2007 测试集上的检测精度和检测速度对比结果

Table 4 Comparison results of detection accuracy and detection speed on PASCAL VOC2007 test set

Method	Dataset	Backbone	Input size	FPS	mAP /%
Faster[4]	VOC07+12	VGG16	600×1000	7	73.2
SSD(Baseline)[10]	VOC07+12	VGG16	300×300	59	74.3
SSD*[10]	VOC07+12	VGG16	300×300	52.6	76.9
DSSD[11]	VOC07+12	ResNet-101	321×321	13.6	78.6
DSOD[29]	VOC07+12	DS/64-192-48-1	300×300	17.4	77.7
RSSD[12]	VOC07+12	VGG16	300×300	35	78.5
FSSD[30]	VOC07+12	VGG16	300×300	**65.8**	78.8
ESSD[31]	VOC07+12	VGG16	300×300	25	79.4
FASSD[32]	VOC07+12	ResNet-50	300×300	30	78.1
DFSSD[33]	VOC07+12	DenseNet-S-32-1	300×300	11.6	78.9
FDSSD[17]	VOC07+12	VGG16	300×300	12.6	79.1
OBSSD	VOC07+12	VGG16	300×300	41.7	**80.8**

SSD* 是在 SSD 的基础上经过数据增强操作的再训练模型。从表 4 可看出，与两阶段目标检测模型 Faster 相比较，OBSSD 模型在检测精度和检测速度两个方面都有了很大的提升。在输入图片尺寸相近的情况下，所提模型与其他模型相比，检测精确度更高：OBSSD 与基准模型 SSD 相比，精确度提升 6.5 个百分点，由于增加了模型参数量，FPS 虽下降 16.5，但也满足 FPS 不低于 30 的实时检测需求；相较 SSD*，所提模型在检测精度方面提升 3.9 个百分点。且在检测速度相差不大的前提条件下，如 FPS 为 35 的 RSSD，所提模型的 mAP 比其高 2.3 个百分点；在检测精确度相差最小的情况下，如 mAP 为 79.4% 的 ESSD，所提模型的 FPS 比其高 16.7。由于独立显卡的不同，检测速度的对比并不绝对。综合来看，所提模型更好地权衡了目标检测的精度和速度。特别地，相较 FPS 值最高的 FSSD，所提模型在损失少许检测速度的情况下有效提升了检测精度，图 7 为 FSSD 模型框架。

与所提模型相比，两者的特征融合方法存在差别，FSSD 模型借鉴特征金字塔网络的思想，重构一组金字塔特征图，即将网络中的 Conv4_3、Conv7 和 Conv8_2 调整为同一尺寸和通道数，然后将特征拼接为一个像素层并进行批归一化处理，再以此层为基础层来生成金字塔特征图，最后进行特征检测。而所提模型基于双向特征融合的思想，引入特征融合权重分层融合 Conv4_3、

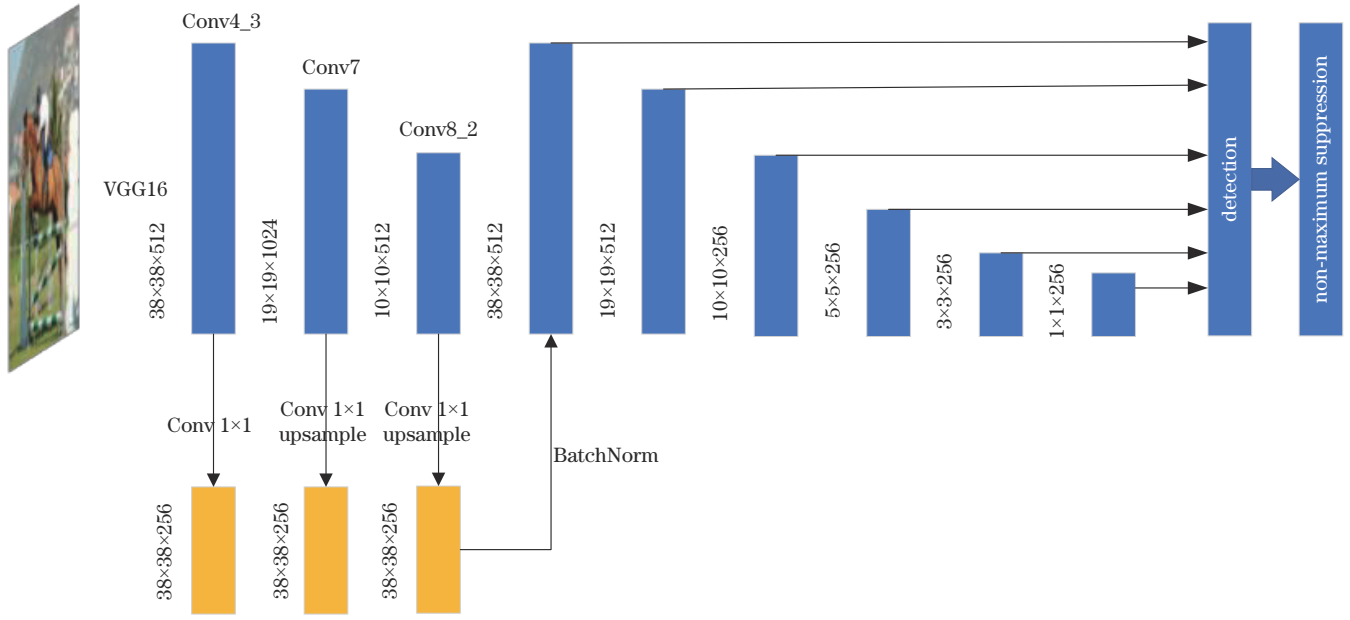

图 7　FSSD 模型框架

Fig. 7　FSSD model framework

Conv7、Conv8_2、Conv9_2、Conv10_2 和 Conv11_2，学习不同特征的重要性。相比 FSSD，所提模型更加充分利用浅层特征中利于目标回归的细节信息和深层特征中利于目标分类的高语义信息。另外，所提模型还在分类和回归预测前引入基于残差模块的检测单元，提高目标定位准确性，有效提升了目标检测精度。

为进一步研究所提模型与部分主流模型在 PASCAL VOC2007 测试集中各类别的检测精度表现情况，对比了在 mAP@0.5（mAP@0.5 表示预测框与真实框的交并比大于等于 0.5 的情况下可以准确预测的概率）的条件下各模型的测试效果。表 5 为具体实验结果。

表 5　PASCAL VOC2007 测试集中 20 个类别平均准确率结果对比

Table 5　Comparison of average precision results of 20 categories in PASCAL VOC2007 test set

Method	mAP /%	areo	bicycle	bird	boat	bottle	bus	car	cat	chair	cow
Faster[4]	73.2	76.5	79.0	70.9	65.5	52.1	83.1	84.7	86.4	52.0	81.9
SSD[10](baseline)	74.3	75.5	80.2	72.3	66.3	47.6	83.0	84.2	86.1	54.7	78.3
SSD*[10]	76.9	76.9	86.6	74.5	66.4	50.4	85.0	84.7	87.3	61.0	78.7
DSSD[11]	78.6	81.9	84.9	80.5	68.4	53.9	85.6	86.2	88.9	61.1	83.5
ESSD[31]	79.4	82.6	86.1	79.8	**72.2**	**54.7**	86.8	86.9	88.2	62.8	85.2
OBSSD	**80.8**	**82.7**	**89.7**	**81.5**	71.8	53.7	**90.7**	**90.0**	**90.6**	**64.8**	**86.2**
Model	mAP /%	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv
Faster[4]	73.2	65.7	84.8	84.6	77.5	76.7	38.8	73.6	73.9	83.0	72.6
SSD[10](baseline)	74.3	73.9	84.5	85.3	82.6	76.2	48.6	73.9	76.0	83.4	74.0
SSD*[10]	76.9	78.2	86.1	89.4	86.0	79.8	48.5	76.1	80.3	86.9	76.1
DSSD[11]	78.6	**78.7**	86.7	88.7	86.7	79.7	51.7	78.0	80.9	87.2	79.4
ESSD[31]	79.4	78.2	87.5	88.0	87.0	80.0	**56.1**	80.2	80.4	88.7	78.1
OBSSD	**80.8**	77.3	**87.9**	**90.0**	**88.1**	**82.0**	54.2	**80.5**	**83.1**	**90.2**	**80.0**

所提模型在大多数类别上的平均准确率达到最高，特别地，在某些类别上的提升效果显著：以 chair 为例，OBSSD 比 SSD 高 10.1 个百分点，较 SSD* 提升 3.8 个百分点；以 bird 为例，OBSSD 相较 SSD 提升 9.2 个百分点，比 SSD* 高 7 个百分点。另外，所提模型在 bus、car、cat、horse 和 train 这 5 个类别的平均准确率达 90% 及以上，较基准模型有着不同程度的提升。为了更直观地观察各模型在 PASCAL VOC2007 测试集上 20 个类别平均准确率的表现情况，现将表 5 的数据可视化，结果如图 8 所示。

由图 8 可知，OBSSD 模型在大多数类别上的平均准确率都比其他模型高。特别地，OBSSD 模型在各类别上的平均准确率较 SSD 模型和 SSD* 模型都有着不同程度的提升，这也证明了所提改进方法的有效性。

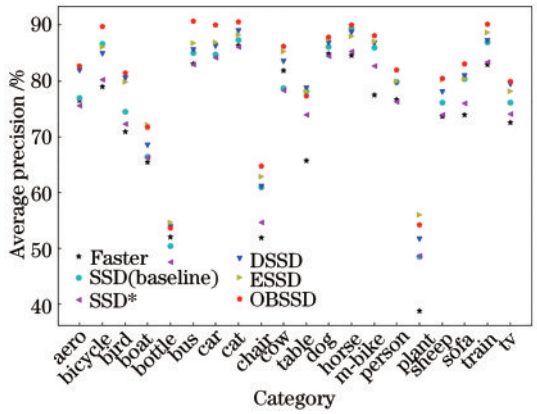

图 8　PASCAL VOC2007测试集上目标检测模型各类别平均准确率对比

Fig. 8　Comparison of average precision of object detection model in PASCAL VOC2007 test set

图 9 展示了所提模型和 SSD* 模型在 PASCAL VOC2007测试集上部分图片的检测结果。

对于每一组图片，左侧的是SSD*模型检测结果，右侧的为OBSSD模型检测结果。从图 9可以看出，SSD*模型存在漏检和误检情况。例如图 9(a)中，SSD*模型将牛错误识别为马，其余图片的检测情况表现为对不能有效识别被遮挡的物体，出现小目标漏检问题。对比检测结果可知，OBSSD模型对于小目标的检测效果提升明显，说明所提单次双向特征金字塔模块通过分层融合，为负责检测小目标的浅层特征提供了丰富的语义信息。同时，对于SSD*模型和OBSSD模型都能检测出的目标，OBSSD模型检测图片上的目标表现出更高的分数，进一步验证了所提模型具备更强的学习能力。

3.3　消融实验

为了对比基准SSD模型、经过数据增强训练得到的SSD*模型和在SSD*模型基础上添加模块所得模型的具体性能，采用控制变量法，设计了相关对照实验。PMSSD*表示在SSD*模型的基础上添加预测模块(PM)的模型，OBMSSD*表示在SSD*模型的基础上添加单次双向特征金字塔模块(OBM)的模型，OBSSD表示在SSD*模型的基础上添加PM与OBM的模型。表 6为各模型在PASCAL VOC2007测试集上具体的实验结果。

表 6的结果表明：以mAP@0.5为例，基准模型SSD的平均精确度为74.3%，经过数据增强后的再训练模型SSD*可以达到76.9%，单独加入PM的PMSSD*模型为78.2%，只添加OBM的OBMSSD*模型达到80.1%，同时添加PM和OBM的OBSSD*模型的平均精确度最大，为80.8%。在SSD*模型的基础上加入所提模块在平均精度上相比基准模型SSD和再训练模型SSD*有着不同程度的提高，由于模型参数量增大和网络结构差异，模型检测速度稍慢于SSD模型，但FPS都在40以上，能够满足实时检测的现实需求。

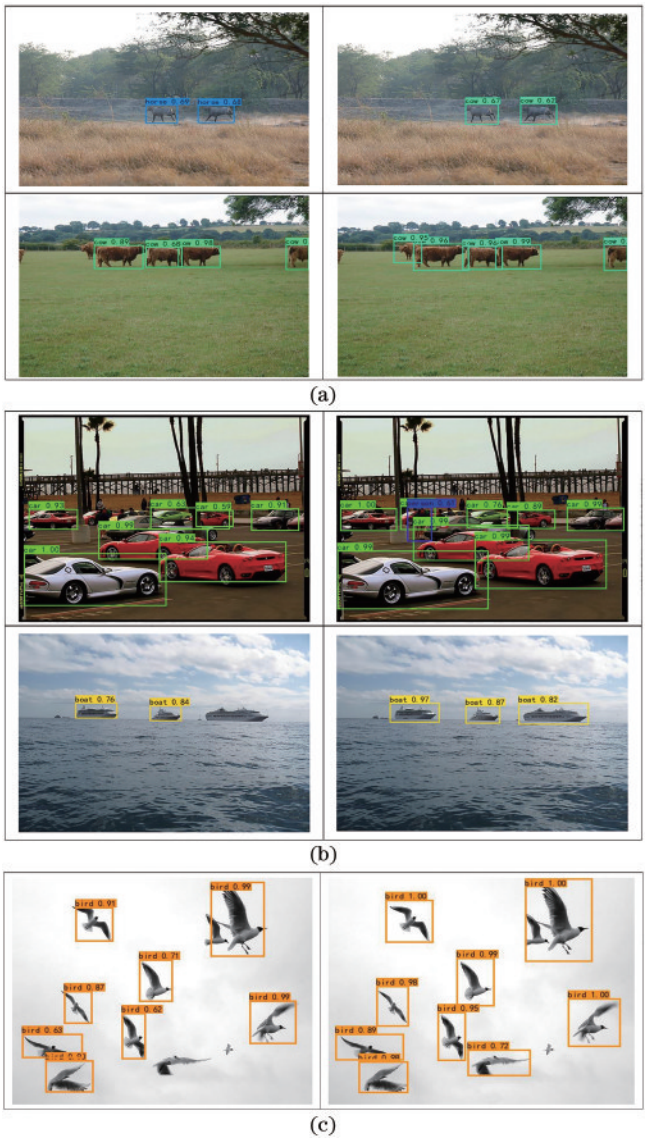

图 9　OBSSD模型与SSD*模型检测结果对比。(a)牛；(b)车、船；(c)鸟、盆栽

Fig. 9　Comparison of detection results between OBSSD model and SSD* model. (a) cow; (b) car, boat; (c) bird, potted plants

表 6　消融实验结果

Table 6　Results of ablation experiment

Model	mAP@0.3 /%	mAP@0.5 /%	Size /MB	FPS
SSD[10]		74.3	25.1	**59**
SSD*[10]	80.8	76.9	25.1	52.6
PMSSD*	82.9	78.2	25.6	48.2
OBMSSD*	84.2	80.1	25.8	44.3
OBSSD*	**85.2**	**80.8**	27.4	41.7

4　结　　论

SSD模型的突出缺点即小目标检测性能低，原因

是SSD模型对浅层特征利用不足，特征提取不充分，浅层特征的低语义信息导致网络不能很好地提取小目标的有用特征，最终降低了小目标检测准确率。所提检测模型旨在加强对浅层特征的利用，引入单次双向特征金字塔模块加强特征提取并有效融合高层特征和浅层特征所含的有用信息；其次，引入融合权重进行特征融合，丰富低层特征的语义信息；此外，在目标分类和回归预测前加入基于残差的预测模块提取更深维度的特征，有效改善浅层特征语义信息不足的问题来提高模型对小目标的检测精确度。随着实时检测的现实需求不断增高，模型精简化成为目标检测研究的一个重要分支。在后续的研究中，可尝试将所提模型的主干网络替换成特征提取能力更加优秀并且网络参数量更低的网络结构，在保证模型检测精度不低于所提模型的情况下，提升模型的检测速度。

参 考 文 献

[1] Voulodimos A, Doulamis N, Doulamis A, et al. Deep learning for computer vision: a brief review[J]. Computational Intelligence and Neuroscience, 2018, 2018: 7068349.

[2] Esteva A, Chou K, Yeung S, et al. Deep learning-enabled medical computer vision[J]. Npj Digital Medicine, 2021, 4: 5.

[3] Zhao Z Q, Zheng P, Xu S T, et al. Object detection with deep learning: a review[J]. IEEE Transactions on Neural Networks and Learning Systems, 2019, 30(11): 3212-3232.

[4] Ren S Q, He K M, 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.

[5] 孙跃军，屈赵燕，李毅红. 基于改进的Mask R-CNN的乳腺肿瘤目标检测研究[J]. 光学学报, 2021, 41(2): 0212004.
Sun Y J, Qu Z Y, Li Y H. Study on target detection of breast tumor based on improved Mask R-CNN[J]. Acta Optica Sinica, 2021, 41(2): 0212004.

[6] He K M, Gkioxari G, Dollár P, et al. Mask R-CNN [C]//2017 IEEE International Conference on Computer Vision, October 22-29, 2017, Venice, Italy. New York: IEEE Press, 2017: 2980-2988.

[7] 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, June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE Press, 2016: 779-788.

[8] 张官荣，陈相，赵玉，等. 面向小目标检测的轻量化YOLOv3算法[J]. 光学学报, 2022, 59(16): 1610008.
Zhang G R, Chen X, Zhao Y, et al. A lightweight YOLOv3 algorithm for small target detection[J]. Acta Optica Sinica, 2022, 59(16): 1610008.

[9] 王建军，魏江，梅少辉，等. 面向遥感图像小目标检测的改进YOLOv3算法[J]. 计算机工程与应用, 2021, 57(20): 133-141.
Wang J J, Wei J, Mei S H, et al. An improved YOLOv3 for small object detection in remote sensing images[J]. Computer Engineering and Applications, 2021, 57(20): 133-141.

[10] Liu W, Anguelov D, Erhan D, et al. SSD: single shot MultiBox detector[M]//Leibe B, Matas J, Sebe N, et al. Computer vision-ECCV 2016. Lecture notes in computer science. Cham: Springer, 2016, 9905: 21-37.

[11] Fu C Y, Liu W, Ranga A, et al. DSSD: deconvolutional single shot detector[EB/OL]. (2017-01-23)[2021-05-06]. https://arxiv.org/abs/1701.06659.

[12] Jeong J, Park H, Kwak N. Enhancement of SSD by concatenating feature maps for object detection[C]// Proceedings of the British Machine Vision Conference 2017, September 4-7, 2017, London, UK. London: British Machine Vision Association, 2017: 76.1-76.12.

[13] 陈幻杰，王琦琦，杨国威，等. 多尺度卷积特征融合的SSD目标检测算法[J]. 计算机科学与探索, 2019, 13(6): 1049-1061.
Chen H J, Wang Q Q, Yang G W, et al. SSD object detection algorithm with multi-scale convolution feature fusion[J]. Journal of Frontiers of Computer Science and Technology, 2019, 13(6): 1049-1061.

[14] 周永福，李文龙，胡冉冉. 多尺度特征融合的双通道SSD行人头部检测算法[J]. 激光与光电子学进展, 2021, 58(24): 2415009.
Zhou Y F, Li W L, Hu R R. Two-channel SSD pedestrian head detection algorithm based on multi-scale feature fusion[J]. Laser & Optoelectronics Progress, 2021, 58(24): 2415009.

[15] 李青援，邓赵红，罗晓清，等. 注意力与跨尺度融合的SSD目标检测算法[J/OL]. 计算机科学与探索: 1-14 [2021-05-30]. http://kns.cnki.net/kcms/detail/11.5602.TP.20210323.1748.013.html.
Li Q Y, Deng Z H, Luo X Q, et al. SSD Object detection algorithm with attention and cross-scale fusion [J/OL]. Journal of Frontiers of Computer Science and Technology: 1-14 [2021-05-30]. http://kns.cnki.net/kcms/detail/11.5602.TP.20210323.1748.013.html.

[16] 郭瑞鸿，张莉，杨莹，等. 基于改进SSD的X光图像管制刀具检测与识别[J]. 激光与光电子学进展, 2021, 58(4): 0404001.
Guo R H, Zhang L, Yang Y, et al. X-ray image controlled knife detection and recognition based on improved SSD[J]. Laser & Optoelectronics Progress, 2021, 58(4): 0404001.

[17] Yin Q J, Yang W Z, Ran M Y, et al. FD-SSD: an improved SSD object detection algorithm based on feature fusion and dilated convolution[J]. Signal Processing: Image Communication, 2021, 98: 116402.

[18] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[EB/OL]. (2014-09-04)[2021-05-04]. https://arxiv.org/abs/1409.1556.

[19] Agarap A F. Deep learning using rectified linear units (ReLU) [EB/OL]. (2018-03-22) [2021-05-06]. https://arxiv.org/abs/1803.08375.

[20] Yu F, Koltun V. Multi-scale context aggregation by

dilated convolutions[EB/OL]. (2015-11-23)[2021-05-04]. https://arxiv.org/abs/1511.07122.

[21] Tan M X, Pang R M, Le Q V. EfficientDet: scalable and efficient object detection[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 13-19, 2020, Seattle, WA, USA. New York: IEEE Press, 2020: 10778-10787.

[22] 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, July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 936-944.

[23] Liu S, Qi L, Qin H F, et al. Path aggregation network for instance segmentation[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE Press, 2018: 8759-8768.

[24] Chollet F. Xception: deep learning with depthwise separable convolutions[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition, July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 1800-1807.

[25] Ramachandran P, Zoph B, Le Q V. Searching for activation functions[EB/OL]. (2017-10-16)[2021-06-04]. https://arxiv.org/abs/1710.05941.

[26] 张震, 李孟洲, 李浩方, 等. 改进SSD算法及其在地铁安检中的应用[J]. 计算机工程, 2021, 47(7): 314-320.
Zhang Z, Li M Z, Li H F, et al. Improved SSD algorithm and its application in subway security detection [J]. Computer Engineering, 2021, 47(7): 314-320.

[27] 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.

[28] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017, 60(6): 84-90.

[29] Shen Z Q, Liu Z, Li J G, et al. DSOD: learning deeply supervised object detectors from scratch[C]//2017 IEEE International Conference on Computer Vision, October 22-29, 2017, Venice, Italy. New York: IEEE Press, 2017: 1937-1945.

[30] Li Z X, Zhou F Q. FSSD: feature fusion single shot multibox detector[EB/OL]. (2017-12-04) [2021-05-04]. https://arxiv.org/abs/1712.00960.

[31] Zheng L W, Fu C M, Zhao Y. Extend the shallow part of single shot multibox detector via convolutional neural network[J]. Proceedings of SPIE, 2018, 10806: 1080613.

[32] Lim J S, Astrid M, Yoon H J, et al. Small object detection using context and attention[C]//2021 International Conference on Artificial Intelligence in Information and Communication (ICAIIC), April 13-16, 2021, Jeju Island, Korea (South). New York: IEEE Press, 2021: 181-186.

[33] Zhai S P, Shang D R, Wang S H, et al. DF-SSD: an improved SSD object detection algorithm based on DenseNet and feature fusion[J]. IEEE Access, 2020, 8: 24344-24357.