DOI: 10.11992/tis.201905041
网络出版地址: http://kns.cnki.net/kcms/detail/23.1538.TP.20190830.1436.002.html

基于跳跃连接金字塔模型的小目标检测

单义[1,2], 杨金福[1,2], 武随烁[1,2], 许兵兵[1,2]
（1. 北京工业大学 信息学部, 北京 100124; 2. 计算智能与智能系统北京重点实验室, 北京 100124）

摘　要: 随着深度学习的发展, 目标检测已经获得了较高的精度和效率。但是小目标的检测仍然是一个挑战。小目标检测准确率较低的重要原因是没有充分利用高层特征的语义信息和低层特征的细节信息之间的关系。针对上述问题, 本文提出一种基于跳跃连接金字塔模型的小目标检测方法。与其他的目标检测方法不同, 本文提出利用跳跃连接金字塔结构来融合多层高层语义特征信息和低层特征图的细节信息。而且为了更好地提取不同尺度物体对应的特征信息, 在网络模型中采用不同大小的卷积核和不同步长的空洞卷积来提取全局特征信息。在 PASCAL VOC 和 MS COCO 数据集上进行了实验, 验证了算法的有效性。

关键词: 跳跃连接金字塔; 全局感受野; 目标检测; 深度学习; 特征提取; 卷积神经网络; 空洞卷积; 图像处理
中图分类号: TP183　　**文献标志码:** A　　**文章编号:** 1673−4785(2019)06−1144−08

中文引用格式: 单义, 杨金福, 武随烁, 等. 基于跳跃连接金字塔模型的小目标检测 [J]. 智能系统学报, 2019, 14(6): 1144–1151.
英文引用格式: SHAN Yi, YANG Jinfu, WU Suishuo, et al. Skip feature pyramid network with a global receptive field for small object detection[J]. CAAI transactions on intelligent systems, 2019, 14(6): 1144–1151.

Skip feature pyramid network with a global receptive field for small object detection

SHAN Yi[1,2], YANG Jinfu[1,2], WU Suishuo[1,2], XU Bingbing[1,2]
(1. Beijing University of Technology, Faculty of Information Technology, Beijing 100124, China; 2. Beijing Key Laboratory of Computational Intelligence and Intelligence System, Beijing 100124, China)

Abstract: With the development of deep learning, objects can be detected with high accuracy and efficiency. However, the detection of small objects remains challenging. The main reason for this is that the relationship between high-level semantic information and low-level feature maps is not fully utilized. To solve this problem, we propose a novel detection framework, called the skip feature pyramid network with a global receptive field, to improve the ability to detect small objects. Unlike previous detection architectures, the skip feature pyramid architecture fuses high-level semantic information with low-level feature maps to obtain detailed information. To extract global information from a network, we apply a global receptive field (GRF) with convolution kernels of different sizes and different dilated convolution steps. The experimental results on PASCAL VOC and MS COCO datasets show that the proposed approach realizes significant improvements over other comparable detection models.

Keywords: skip feature pyramid network; global receptive field; object detection; deep learning; feature extraction; convolutional neural network; dilated convolution; image processing

目标检测是实现场景理解的基础, 是计算机视觉领域的一项重点研究内容。自从深度卷积网络大幅度提高图像分类的准确率之后, 深度卷积神经网络也广泛应用于目标检测中。虽然深度卷积神经网络在目标检测方面取得了巨大的进步, 但是小目标的检测仍存在检测准确率较低的问题。

现阶段基于深度学习的目标检测算法主要分为两类, 一是基于分类的目标检测方法。基于分类的目标检测算法又称为两阶段 (two-stage) 模型, 首先选取候选区域, 然后对候选区域进行分类和位置回归, 最终输出检测的结果。2014 年

收稿日期: 2019−05−23.　**网络出版日期:** 2019−08−30.
基金项目: 国家自然科学基金项目 (6153302); 北京市自然科学基金项目 (4182009).
通信作者: 单义. E-mail: 15732036708@163.com.

Girshick 等[1] 首次提出基于区域提取的 R-CNN 算法。2015 年 Girshick[2] 又提出一种改进的 Fast R-CNN 算法, 将图像经过基础网络处理之后, 再传入 R-CNN 子网络, 共享卷积运算。但 Fast R-CNN 在提取区域候选框时仍然使用选择性搜索算法 (Selective Search[3]), 增加了算法耗时, 运行速度慢。针对 Fast R-CNN 算法的缺点, 2015 年 Ren 等[4] 提出 Faster R-CNN 算法, 用候选区域生成网络 (regional proposal networks, RPN) 来代替选择性搜索算法。另一类是基于回归的目标检测算法, 又称为一阶段 (one-stage) 模型。2016 年 Redmon 等[5] 提出了新的目标检测算法 YOLO(You Only Look Once)。YOLO 算法将目标检测框架看作空间上的回归问题。但是 YOLO 算法存在定位精度、召回率等较低的问题, 且对尺寸较小的物体检测效果不好, 泛化能力相对较弱。为了解决 YOLO 算法的缺陷, 2016 年 Liu 等[6] 提出 SSD (single shot multiBox detector) 算法, 利用多层特征图进行检测。

此外, 针对小目标的物体检测存在的问题, 有学者提出了新的检测网络模型。2016 年 Bell 等[7] 提出一种利用感兴趣区域内外信息进行物体检测的模型 (inside-outside net, ION)。2017 年 Lin 等[8] 在 Faster R-CNN 网络的基础上提出一种具有横向连接的特征金字塔网络 (feature pyramid networks, FPN),利用多尺度特征和自上而下的结构实现目标检测。FPN 只利用顶层的特征进行检测, 虽然信息丰富, 但是经过层层池化, 很多细节特征信息会丢失, 而这些信息对小目标检测具有重要意义。Fu 等[9] 针对 SSD 算法在小目标检测上存在的问题, 提出一种改进的 DSSD(deconvolutional single shot detector) 算法, 将 SSD 算法的基础网络更改为 ResNet-101[10], 增强了网络的特征提取能力, 结合多尺度信息, 提高了检测结果。然而, 上述网络忽略了低层特征与高层特征之间的联系, 并且对于不同尺度大小的物体, 卷积操作的感受野是不同的, 在基础网络中利用同一种大小的卷积核进行卷积运算, 不能很好提取不同大小的物体的感受野信息。

针对上述问题, 本文提出基于跳跃连接金字塔和全局感受野的网络结构, 来融合不同尺度的高层与底层的特征信息, 并利用不同大小的卷积核和不同步长的空洞卷积[11] 来提取全局特征信息。实验结果表明, 所提出的模型能有效改善小目标检测结果。

1　算法模型

图 1 是本文提出的检测模型的整体结构图。网络模型基于前馈深度卷积网络, 通过跳跃连接的金字塔 (skip feature pyramid network)结构。将高层特征与低层特征相融合, 在基础网络中加入全局感受野模块 (global receptive field, GRF) 来提取不同尺度的物体的全局特征信息。该网络模型在所有特征层中选取 4 层不同尺度的高层与低层特征图进行预测, 在预测层产生固定数量的包围盒 (anchor boxes) 和相应的类别概率值, 最后通过非极大值抑制 (non-maximum suppression) 来获得最终的预测框。

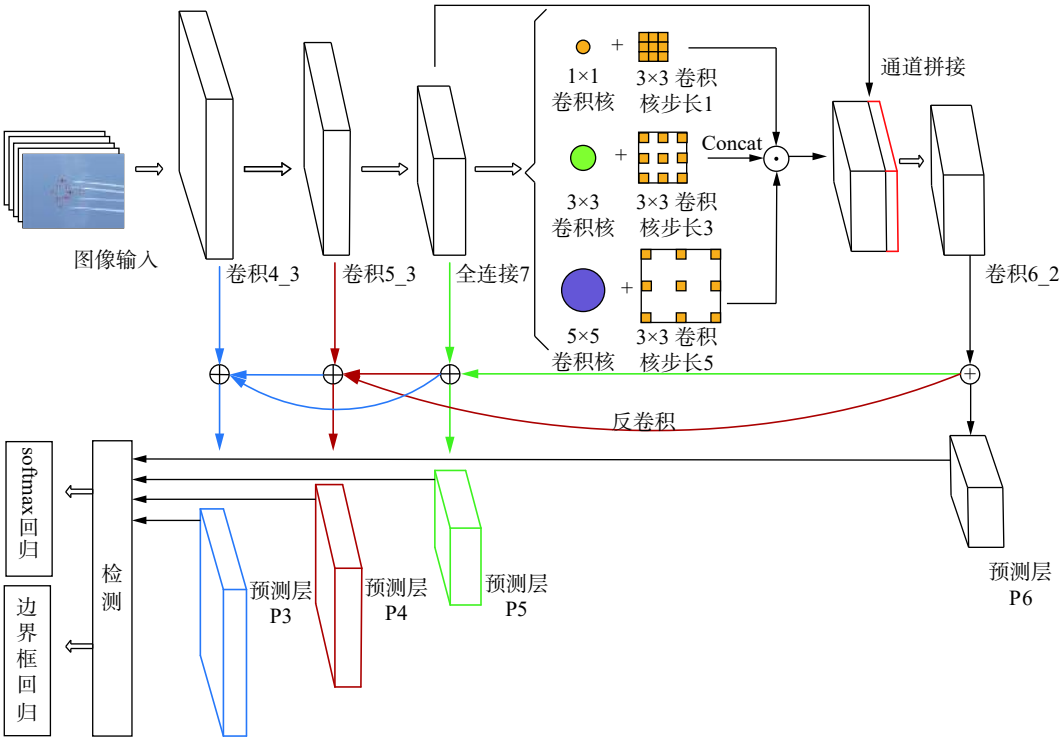


图 1　基于跳跃连接金字塔的小目标检测模型

Fig. 1　Title Skip feature pyramid network with global receptive field for object detection

为了能够清晰地展示本文提出的网络结构，只展示用于预测层的特征图，采用不同的颜色表示不同的特征层之间的融合，构建跳跃连接金字塔结构，并且提出利用全局感受野模块来提取网络全局特征信息。

1.1　跳跃连接金字塔

如图 2 所示，(a)~(c) 为现阶段应用较多的深度网络检测结构。图 2(d) 为本文提出的跳跃连接金字塔网络结构。图 2(a) 中只利用深度网络中最后一层特征图进行预测，其中 YOLO 算法就是采用图 2(a) 的结构形式，该网络具有很高的检测速度，但是检测的准确率较低。图 2(b) 是对图 2(a) 算法结构的改进，通过在不同的尺度的高层特征上预测，有效地弥补图 2(a) 结构中存在的问题，改善了检测结果。图 2(c) 是一种自顶而下的金字塔结构模型，不仅采用不同的特征层进行预测，而且融合了相邻特征图之间的信息。但是这种结构的网络模型忽略了不同高层特征图与低层特征图之间的联系。针对上述问题，提出如图 2(d) 所示的跳跃金字塔结构模型，采用跳跃连接的方式，通过选择不同步长的反卷积进行上采样高层特征图，并使用逐像素求和的计算方法来融合不相邻的特征图之间的信息。

在深度卷积网络中，网络的最深层的特征图包含最多的抽象特征信息。因此利用提出跳跃金字塔结构，来融合不同的高层和低层特征图之间的信息，不仅能有效利用不同特征层之间的尺度信息，而且融合了高层特征图与低层特征图之间的细节信息。

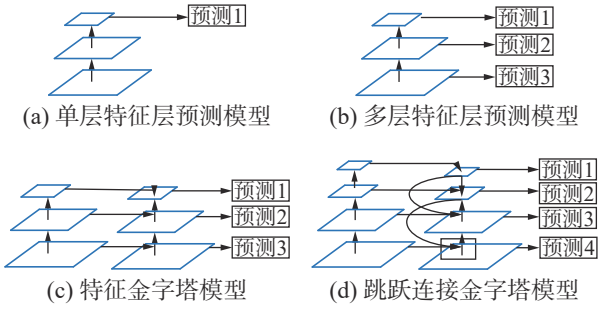


图 2　不同的结构形式的预测网络

Fig. 2　Different structure of predicting network

跳跃连接的金字塔的细节结构如图 3 所示。通过选择基础网络中的高层特征图，对每一个高层特征图利用大小为 3×3 通道数为 256 卷积核进行卷积操作。这样做的目的是将不同特征层的特征图的通道数变成相同的数量，以便进行融合计算。在统一了每一层的通道数之后，采用 2×2 步长为 2 的反卷积操作来把相邻特征层的特征图进行上采用，上采样之后不同特征层就变为相同的大小。利用 4×4 步长为 4 的反卷积来上采样不相邻的特征图。反卷积计算特征图的大小计算公式为

$$o=[\frac{i-f+2p}{s}]+1 \tag{1}$$

式中：i 示输入特征图的尺寸；f 为卷积核的大小；p 为填充的像素数；s 为反卷积的步长。最后对不同特征层不同通道数的特征层，可以用每个像素对应点的和来作为融合之后的特征图。

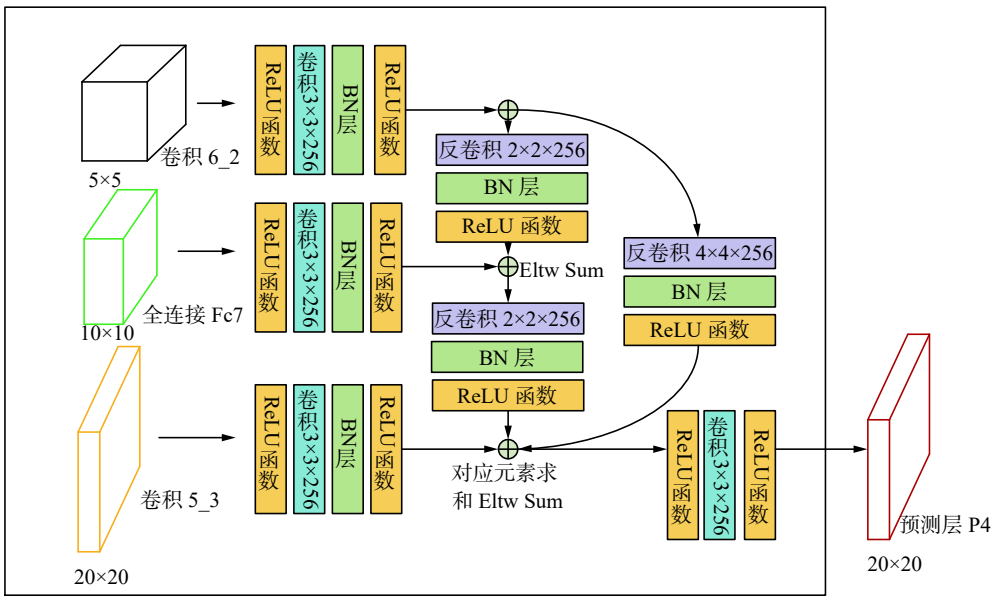


图 3　跳跃连接的金字塔的细节结构

Fig. 3　The detailed structure of skip feature pyramid

1.2　全局感受野模块

在大多数的检测模型中, 多采用自上而下的结构, 忽略了对于不同的大小的物体。而卷积神经网络的感受野是不同的, 因此提出在网络中加入横向连接的结构, 采用不同大小的卷积核和不同步长的空洞卷积来增强网络全局的感受野。与只利用一种卷积核的网络结构相比, 使用不同的大小的卷积核和不同步长的空洞卷积能有效提取不同尺度大小物体的特征。首先, 利用1×1的卷积层, 改变特征图的通道数, 减少特征模型的计算量。然后, 利用 1×1、3×3 和 5×5 三种不同的卷积核和 3×3 步长分别为 1, 3 和 5 的空洞卷积[11]来提取不同尺度的特征信息。进而把获取的特征通道进行连接, 用 1×1 的卷积核将通道数变换为原来特征图相同的通道数, 并与原来的特征图对应像素点进行叠加, 既保留原本网络模型的特征, 又加入了不同大小感受野的特征信息, 有效改善提取较小尺寸的物体特征。全局感受野结构如图 4 所示。

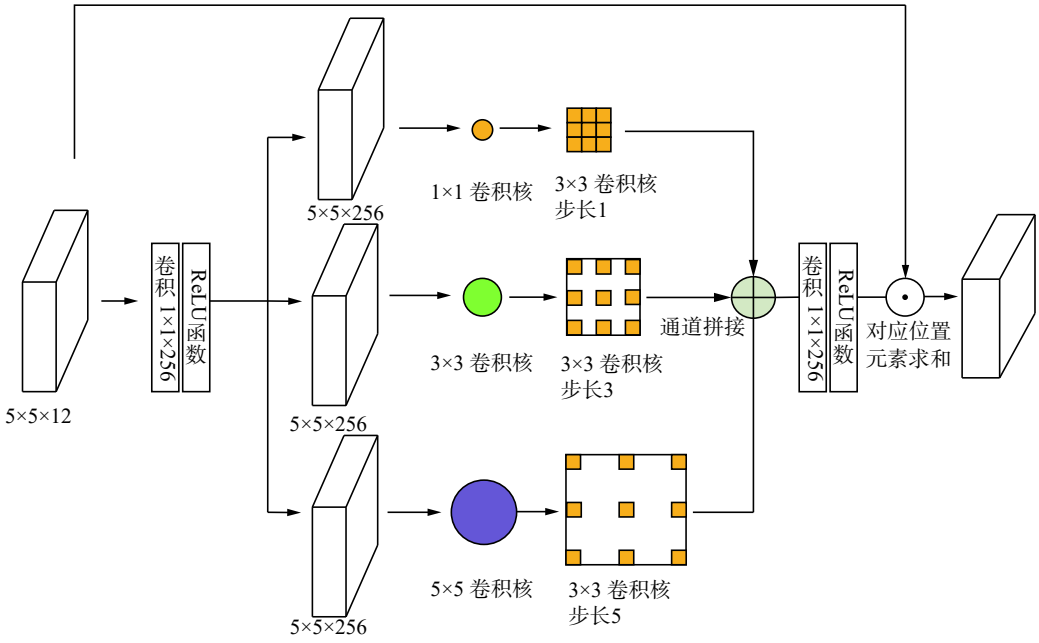


图 4　全局感受野结构

Fig. 4　The network of global receptive field

1.3　包围框的设置

为了更好地利用图像的特征信息, 选取 VGG16 中 4 层不同尺度的特征图, 每一层对应的步长分别为 8、16、32、64 个像素, 在特征图的每个像素点采用不同大小比例的包围盒进行预测。每一层特征图, 设置不同大小的比例包围盒来进行预测不同大小的物体, 纵横比分别为 0.5、1.0、2.0。在训练时, 当包围盒与图像标注 (ground truth) 的重叠面积比例大于 0.5 时, 即认为该包围盒中存在目标物体。

1.4　损失函数

考虑到正负样本的数量相差较大, 与 SSD 类似, 本文采用负挖掘来解决极端的前景背景类别不平衡的问题。即在训练中, 不使用所有的负样本包围盒, 也不随机选择负样本包围盒, 而是将负样本的损失进行排序, 选择其中损失最大的负样本作为最后预测的样本, 并且控制最终的正负样本比例为 3:1。与 SSD 不同的是, 本文在进行预测之前, 先对网络预测产生的包围盒进行前景和背景的二分类滤除。这样做的目的在于有效减少负样本的数量。网络的损失函数为

$$L(\{p_i\},\{x_i\},\{c_i\},\{t_i\})=\frac{1}{N_{\text{conv}}}\left(\sum_i l_b(p_i,[l_i^*\geqslant 1])+\sum_i [l_i^*\geqslant 1]\right.$$
$$\left. l_r(x_i,g_i^*)\right)+\frac{1}{N_p}\left(\sum_i l_m(c_i,l_i^*)+\sum_i [l_i^*\geqslant 1] l_r(t_i,g_i^*)\right) \tag{2}$$

式中: i 是每个训练批次中包围盒的索引; l_i^* 是每个批次图像中每个图像标注的对应类别标签; g_i^* 是每个图像标注对应的坐标; p_i 和 x_i 是网络预测的包围盒中是否有目标和相应的坐标信息; c_i 和 t_i 是所预测的目标包围盒中物体的类别和相应的坐标信息; N_{conv} 和 N_p 分别为特征提取网络和预测网络中正样本包围盒的数量; l_b 是特征提取网络输出的二分类的交叉熵损失, 即判断包围盒中是否有目标; l_m 是多分类任务的置信度。与 Fast R-CNN 算法相似, l_r 为 smooth L1 回归损失。只有当包围盒中 $l_i^*\geqslant 1$ 时, 即预测值为真时才会计算相应的损失。其中位置损失函数 l_r 的具体损失函数如下:

$$l_r(x,g^*,l^*)=\sum_{i\in \mathrm{Pos}}^{N}\sum_{m\in\{cx,cy,w,h\}}[l^*\geqslant 1]\mathrm{smooth}_{\mathrm{L1}}(x_i^m-\widehat{g}_j^m) \quad (3)$$

$$\widehat{g}_j^{cx}=(g_j^{cx}-d_i^{cx})/d_i^w,\quad \widehat{g}_j^{cy}=(g_j^{cy}-d_i^{cy})/d_i^h \quad (4)$$

$$\widehat{g}_j^{w}=\log(\frac{g_j^w}{d_i^w}),\quad \widehat{g}_j^{h}=\log(\frac{g_j^h}{d_i^h}) \quad (5)$$

其中：cx、cy、w、h分别表示包围盒的中心坐标以及宽和高。(g^{cx},g^{cy},g^w,g^h)表示图像标注信息中的包围盒对应的中心坐标以及宽和高，(d^{cx},d^{cy},d^w,d^h)表示默认包围盒的中心坐标以及宽和高，(x^{cx},x^{cy},x^w,x^h)表示预测的包围盒的中心坐标以及宽和高。

2　实验结果及分析

本文以 VGG16[12] 作为基础的特征提取网络，并且在 ILSVRC CLS-LOC 数据集上进行了预训练。为了验证所提出算法的有效性，分别在 PASCAL VOC[13] 和 MS COCO[14] 数据集上进行了实验。PASCAL VOC 和 MS COCO 数据集中类别数量分别为 20 和 80，并且每个类别都有标注信息和对应的类别信息。

2.1　PASCAL VOC

在 PASCAL VOC 数据上，所有的算法模型都在 VOC2007 和 VOC2012 数据集上进行训练，测试在 VOC2007 数据集上进行测试。一共训练了 140 k 次，学习率不断衰减，设置 0~80 k 的学习率为 10^{-3}，80 k 到 100 k 的学习率下降为 10^{-4}，100 k 到 120 k 学习率下降为 10^{-5}，120k~140 k 的学习率为 10^{-6}。考虑到 GPU 的处理能力，设置不同步长的学习批次，对于输入大小为 320×320 的图像，处理批次设置为 32，而对于输入图像 512×512，设置的批次为 16。动量和权值衰减分别设置为 0.9 和 0.000 5。

表 1 为本文的实验结果和其他网络检测结果的对比。输入图像的尺寸对模型的输出结果有较大影响。从表中可以看出，在输入图像的尺寸为 320×320 时，平均准确率为 80.1%，速度为 31.2 f/s。在输入图像的尺寸为 512×512 时，平均准确率为 81.9%，速度为 18.2 f/s。

表 1　PASCAL VOC2007 的不同网络模型的检测结果

Table 1　Detection results on PASCAL VOC dataset

方法	基础网络	准确率	检测速度	图像尺寸
Faster[4]	VGG16	73.2	7	1 000×600
Faster[4]	Residual-101[10]	76.4	2.4	1 000×600
R-FCN[15]	Residual-101	80.5	9	1 000×600
DSOD300[16]	DS/64-192-4	77.7	17.4	300×300
YOLOv2[17]	Darknet-19	78.6	40	544×544
SSD300[6]	VGG16	77.5	46	300×300
DSSD321[9]	Residual-101	79.5	9.5	321×321
STDN321[18]	DenseNet-169	79.2	41.5	321×321
Ours320	**VGG16**	**80.1**	**31.2**	**320×320**
SSD512[6]	VGG16	78.6	19	512×512
DSSD513[9]	Residual-101	81.5	5.5	513×513
STDN513[18]	DenseNet-169	80.9	28.6	513×513
Ours512	**VGG16**	**81.9**	**18.2**	**512×512**

表 2 为网络模型在 PASCAL VOC2007 测试集的不同类别平均准确率的结果。从表中可以看出，本文方法在小目标类别中的平均检测准确率明显高于其他网络模型。本文所提出的网络整体的平均准确率高于其他网络 1%，其中如 bird、sheep、plant 等小目标比其他网络最优准确率分别高 2.5%、3.2%、2.7%，证明了提出网络的有效性。

表 2　PASCAL VOC2007 不同类别的检测结果

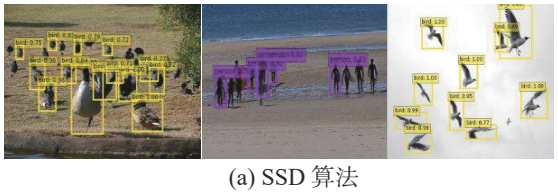
Table 2　Object detection results on PASCAL VOC 2007 test set

类别 方法	Faster[4]	ION[7]	MR-CNN[19]	YOLOv2[17]	SSD 300[6]	SSD 512[6]	STDN 321[18]	STDN 513[18]	**Ours 320**	**Ours 512**
aero	76.5	79.2	80.3	86.3	79.5	84.8	81.2	86.1	84.5	**88.5**
bike	79.0	83.1	84.1	82.0	83.9	85.1	88.3	**89.3**	85.4	86.4
bird	70.9	77.6	78.5	74.8	76.0	81.5	78.1	79.5	80.1	**84.0**
boat	66.5	65.6	70.8	59.2	69.6	73.0	72.2	74.3	73.8	**75.8**
bottle	52.1	54.9	68.5	51.8	50.5	57.8	54.3	61.9	60.0	**69.4**
bus	83.1	85.4	88.0	79.8	87.0	87.8	87.6	88.5	87.7	**88.9**
car	84.7	85.1	85.9	76.5	85.7	88.3	86.7	88.3	88.2	**89.2**
cat	86.4	87.0	87.8	**90.6**	88.1	87.4	88.7	89.4	89.0	89.5
chair	52.0	54.4	60.3	52.1	60.3	63.5	63.5	**67.4**	63.8	66.7
cow	81.9	80.6	85.2	78.2	81.5	85.4	83.2	85.5	84.7	**86.4**
table	65.7	73.8	73.7	58.5	77.0	73.2	79.4	**79.5**	77.2	73.2
dog	84.8	85.3	87.2	**89.3**	86.1	86.2	86.1	86.4	86.0	87.6
horse	84.6	82.2	86.5	82.5	87.5	86.7	**89.3**	89.2	86.4	88.2
mbike	77.5	82.2	85.0	83.4	83.9	83.9	88.0	**88.5**	86.7	87.5
person	76.7	74.4	76.4	81.3	79.4	82.5	77.3	79.3	82.5	**84.9**
plant	38.8	47.1	48.5	49.1	52.3	55.6	52.5	53.0	56.1	**58.3**
sheep	73.6	75.8	76.3	77.2	77.9	81.7	80.3	77.9	81.3	**84.9**
sofa	73.9	72.7	75.5	62.4	79.5	79.0	80.8	**81.4**	80.4	78.3
train	83.0	84.2	85.0	83.4	87.6	86.6	86.3	86.6	88.5	**87.8**
tv	72.6	80.4	81.0	68.7	76.8	80.0	82.1	**85.5**	79.8	80.8
mAP	73.2	75.6	78.2	76.8	77.5	79.5	79.3	80.9	80.1	**81.9**

图 5 为本文的方法和 SSD 方法在 VOC2007 数据集的可视化结果对比图。第一行是 SSD 算法的检测结果, 第二行是本文提出的检测方法的实验结果。从图中可以看出, 本文方法对于图像中较小尺寸的鸟和人的检测效果明显改善, 而且对于正确检测的物体的置信度也有了较大的提高。

2.2　MS COCO

为了进一步验证本文提出的模型在更多类别、更多数量的数据集上的有效性, 我们在 MS COCO 数据集上进行了实验, 实验结果如表 3 所示。MS COCO 数据集的评价指标不同于 PASCAL VOC。以不同的 IOU 进行评价, 对图像分为 3 个规模大小进行评价。其中 AP 表示准确率, AR 表示召回率。APs 和ARs 分别表示小目标的检测准确率和召回率, 以 320×320 尺寸的图像为例, 本文方法在小目标准确率和召回率分别高于其他最优的模型 2.4% 和 4%。



(a) SSD 算法

(b) 本文方法

图 5　在 VOC2007 上可视化的实验结果对比

Fig. 5　The visual comparison of experimental results on VOC2007 test

表 3 MS COCO 数据集检测结果

Table 3 Object detection results on MS COCO test-dev set.

方法	基础网络	$AP_{0.5:0.95}$	$AP_{0.5}$	$AP_{0.75}$	AP_S	AP_M	AP_L	AR_1	AR_{10}	AR_{100}	AR_S	AR_M	AR_L
Faster[4]	VGG16	21.9	42.7	–	–	–	–	–	–	–	–	–	–
ION[7]	VGG16	23.6	43.2	23.6	6.4	24.1	38.3	23.2	32.7	33.2	10.1	37.7	53.6
R-FCN[15]	Residual-101	29.2	51.5	–	10.3	32.4	43.3	–	–	–	–	–	–
DSOD[16]	DS/64/192/4	29.3	47.3	30.6	9.4	31.5	47	27.3	40.7	43	16.7	47.1	65
Yolov2[17]	Darknet	21.6	44.0	19.2	9.0	28.9	41.9	24.8	37.5	39.8	14.0	43.5	59.0
SSD300[6]	VGG16	25.1	43.1	25.8	6.6	25.9	41.4	23.7	35.1	37.2	11.2	40.4	58.4
DSSD321[9]	Residual-101	28.0	46.1	29.2	7.4	28.1	47.6	25.5	37.1	39.4	12.7	42	62.6
STDN321[18]	DenseNet[20]	28.0	45.6	29.4	7.9	29.7	45.1	24.4	36.1	38.4	12.5	42.7	60.1
Ours320	**VGG16**	**28.2**	**47.7**	**29.1**	**10.3**	**31.4**	**43.7**	**25.8**	**38.9**	**41.2**	**16.9**	**47.2**	**61.0**
SSD512[6]	VGG16	28.8	48.5	30.3	10.9	31.8	43.5	26.1	39.5	42	16.5	46.6	60.8
DSSD513[9]	Residual-101	33.2	53.3	35.2	13	35.4	51.1	28.9	43.5	46.2	21.8	49.1	66.4
STDN513[18]	DenseNet	31.8	51.0	33.6	14.4	36.1	43.4	27.0	40.1	41.9	18.3	48.3	57.3
Ours512	**VGG16**	**33.1**	**52.3**	**32.4**	**15.6**	**34.6**	**42.7**	**28.3**	**42.6**	**45.6**	**25.9**	**50.8**	**60.1**

3 结束语

针对小目标检测准确率较低的问题，本文提出了一种基于跳跃连接金字塔的小目标检测模型。通过跳跃连接的特征金字塔融合高层与低层特征图信息，并且利用不同大小卷积和不同步长空洞卷积的横向结构来提取全局特征信息，有效弥补因连续池化而造成的信息丢失。整个网络模型以端到端方式进行训练，并且在 PASCAL VOC 和 MS COCO 数据集上进行了实验，实验结果表明本文提出的模型在小目标的检测准确率方面明显优于其他算法模型。

参考文献：

[1] GIRSHICK R, DONAHUE J, DARRELL T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Columbus, OH, USA, 2014: 580–587.

[2] GIRSHICK R. Fast R-CNN[C]//Proceedings of 2015 IEEE International Conference on Computer Vision (ICCV). Santiago, Chile, 2015: 1440–1448.

[3] UIJLINGS J R R, VAN DE SANDE K E A, GEVERS T, et al. Selective search for object recognition[J]. International journal of computer vision, 2013, 104(2): 154–171.

[4] 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.

[5] REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: unified, real-time object detection[C]//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA, 2016: 779–788.

[6] LIU Wei, ANGUELOV D, ERHAN D, et al. SSD: single shot MultiBox detector[C]//Proceedings of the 14th European Conference on Computer Vision. Amsterdam, the Netherlands, 2016: 21–37.

[7] 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 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA, 2016: 2874–2883.

[8] LIN T Y, DOLLÁR P, GIRSHICK R, et al. Feature pyramid networks for object detection[C]//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI, USA, 2017: 936–944.

[9] FU Chengyang, LIU Wei, RANGA A, et al. DSSD: deconvolutional single shot detector[J]. arXiv: 1701.06659, 2017.

[10] HE Kaiming, ZHANG Xiangyu, REN Shaoqing, et al. Deep residual learning for image recognition[C]//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA, 2016: 770–778.

[11] YU F, KOLTUN V. Multi-scale context aggregation by dilated convolutions[J]. arXiv:1511.07122, 2015.

[12] SIMONYAN K, ZISSERMAN A. Very Deep Convolu-

tional Networks for Large-Scale Image Recognition[J]. arXiv:1409.1556, 2014.

[13] 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.

[14] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft COCO: common objects in context[C]//Proceedings of the 13th European Conference on Computer Vision. Zurich, Switzerland, 2014: 740–755.

[15] DAI Jifeng, LI Yi, HE Kaiming, et al. R-FCN: object detection via region-based fully convolutional networks[C]// Proceedings of the 30th International Conference on Neural Information Processing Systems. Barcelona, Spain, 2016: 379–387.

[16] SHEN Zhiqiang, LIU Zhuang, LI Jianguo, et al. DSOD: learning deeply supervised object detectors from scratch[C]//Proceedings of 2017 IEEE International Conference on Computer Vision (ICCV). Venice, Italy, 2017: 1937–1945.

[17] REDMON J, FARHADI A. YOLO9000: better, faster, stronger[C]//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI, USA, 2017: 6517–6525.

[18] ZHOU Peng, NI Bingbing, GENG Cong, et al. Scale-transferrable object detection[C]//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA, 2018: 528–537.

[19] GIDARIS S, KOMODAKIS N. Object detection via a multi-region and semantic segmentation-aware CNN model[C]//Proceedings of 2015 IEEE International Conference on Computer Vision (ICCV). Santiago, Chile, 2015: 1134–1142.

[20] HUANG Gao, LIU Zhuang, VAN DER MAATEN L, et al. Densely connected convolutional networks[C]//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI, USA, 2017: 2261–2269.

作者简介：

单义，男，1992 年生，硕士研究生，主要研究方向为深度学习、计算机视觉。

杨金福，男，1977 年生，教授，主要研究方向为机器学习、机器视觉、智能计算与智能系统。近年来承担包括国家大科学工程、国家重点研发计划、国家 973 计划、国家 863 计划、国家自然科学基金、北京市自然科学基金等 20 多项科研项目。申请国家发明专利 30 余项 (获得授权 20 余项)，获得软件著作权登记 10 余项，发表学术论文 80 余篇。

武随烁，男，1997 年生，硕士研究生，主要研究方向为深度学习、计算机视觉。