智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

自适应上下文特征的多尺度目标检测算法

王凤随, 陈金刚, 王启胜, 刘芙蓉

引用本文:
王凤随,陈金刚,王启胜,刘芙蓉. 自适应上下文特征的多尺度目标检测算法[J]. 智能系统学报, 2022, 17(2): 276–285.
WANG Fengsui,CHEN Jingang,WANG Qisheng,LIU Furong. Multi–scale target detection algorithm based on adaptive context features[J]. *CAAI Transactions on Intelligent Systems*, 2022, 17(2): 276–285.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202101029

DOI: 10.11992/tis.202101029
网络出版地址: https://kns.cnki.net/kcms/detail/23.1538.TP.20210622.1127.008.html

自适应上下文特征的多尺度目标检测算法

王凤随[1,2,3], 陈金刚[1,2,3], 王启胜[1,2,3], 刘芙蓉[1,2,3]

（1. 安徽工程大学 电气工程学院, 安徽 芜湖 241000; 2. 检测技术与节能装置安徽省重点实验室, 安徽 芜湖 241000; 3. 高端装备先进感知与智能控制教育部重点实验室, 安徽 芜湖 241000）

摘　要: 识别多尺度目标是检测任务中的一项挑战, 针对检测中的多尺度问题, 提出自适应上下文特征的多尺度目标检测算法。针对不同尺度的目标需要不同大小感受野特征进行识别的问题, 构建了一种多感受野特征提取网络, 通过多分支并行空洞卷积, 从高层语义特征中挖掘标签中的上下文信息; 针对不同尺度目标的语义特征出现在不同分辨率特征图中的问题, 基于改进的通道注意力机制, 提出自适应的特征融合网络, 通过学习不同分辨率特征图之间的相关性, 在全局语义特征中融合局部位置特征; 利用不同尺度的特征图识别不同尺度的物体。在 PASCAL VOC 数据集上对本文算法进行验证, 本文方法的检测精度达到了 85.74%, 相较于 Faster R-CNN 检测精度提升约 8.7%, 相较于基线检测算法 YOLOv3+提升约 2.06%。

关键词: 机器视觉; 目标检测; 卷积神经网络; 通道注意力; 并行空洞卷积; 多尺度特征融合; 上下文特征; 深度学习

中图分类号: TP391.4　　**文献标志码**: A　　**文章编号**: 1673−4785(2022)02−0276−10

中文引用格式: 王凤随, 陈金刚, 王启胜, 等. 自适应上下文特征的多尺度目标检测算法 [J]. 智能系统学报, 2022, 17(2): 276–285.

英文引用格式: WANG Fengsui, CHEN Jingang, WANG Qisheng, et al. Multi-scale target detection algorithm based on adaptive context features[J]. CAAI transactions on intelligent systems, 2022, 17(2): 276–285.

Multi-scale target detection algorithm based on adaptive context features

WANG Fengsui[1,2,3], CHEN Jingang[1,2,3], WANG Qisheng[1,2,3], LIU Furong[1,2,3]

(1. School of Electrical Engineering, Anhui Polytechnic University, Wuhu 241000, China; 2. Anhui Key Laboratory of Detection Technology and Energy Saving Devices, Wuhu 241000, China; 3. Key Laboratory of Advanced Perception and Intelligent Control of High-end Equipment, Ministry of Education, Wuhu 241000, China)

Abstract: Multi-scale target recognition is a challenge in any detection task. Aiming at the multi-scale problem in detection, a multi-scale target detection algorithm with adaptive context features is proposed. A multi-receptive field feature extraction network was constructed to solve the problem wherein targets of different scales require different receptive field features to be recognized. Using multi-branch parallel void convolution, contextual information in tags was extracted from high-level semantic features. Based on an improved channel attention mechanism, an adaptive feature fusion network was proposed to solve the problem wherein the semantic features of different scale targets appear in feature maps of different resolutions. The local location features were fused into global semantic features by learning the correlation between feature maps of different resolutions. The feature maps of different scales were used to identify objects of different scales. The proposed algorithm was verified on a Pascal VOC data set; the detection accuracy of the proposed method reached 85.74%, which was approximately 8.7% higher than the Faster R-CNN and about 2.06% higher than the baseline detection algorithm YOLOV3 +.

Keywords: machine vision; target detection; convolution neural network; channel attention; parallel empty convolution; multi-scale feature fusion; contextual feature; deep learning

目标检测作为计算机视觉的一个分支, 随着深度学习模型与检测任务的结合以及 GPU 计算能力的提升, 它在学术和工业界得到广泛的研究和应用, 如人脸识别、行人检测、自动驾驶等领域。

收稿日期: 2021−01−19.　**网络出版日期**: 2021−06−22.
基金项目: 安徽高校省级自然科学研究重点项目 (KJ2019A0162); 安徽省自然科学基金项目 (2108085MF197, 1708085MF154); 检测技术与节能装置安徽省重点实验室开放基金项目 (DTESD2020B02).
通信作者: 王凤随. E-mail: fswang@ahpu.edu.cn.

目标检测领域, 尺度的变化问题一直是个挑战, 它直接影响着检测精度。在检测任务中, 数据集中目标的尺度范围变化较大。小尺度目标经过卷积神经网络后, 由于感受野的增大造成特征的丢失。因此, 神经网络需对不同尺度的目标都可以很好地提取特征。Faster-RCNN[1-2] 作为两阶段目标检测算法的大成之作, 它提出的使用区域建议网络代替选择性搜索 (selective-search) 提取候选框, 多尺度锚框的使用减少了选取候选框的时间, 取得更高的精度和更短的训练时间。但 Faster-RCNN (faster region convolutional neural network) 只利用神经网络的最后一层特征进行预测, 缺乏处理多尺度目标的能力。针对多尺度目标需要多尺度特征预测的问题, SNIP[3-4](scale normalization for image pyramids) 使用图像金字塔将原始图直接进行不同尺度的缩放变化, 利用不同分辨率的图片来检测不同尺度的物体, 在尺度变化的问题上取得不错的成效。这种尺度变化的方式虽然有效, 但也大大增加了检测的时间复杂度。另一种尺度变化的思想是利用特征金字塔 (feature pyramid) 来近似图像金字塔, FPN[5](feature pyramid network) 利用对高层语义信息上采样, 以自上而下的方式增强低层特征, FPN 在 YOLOv3[6] 中具有很好的多尺度表现。但 FPN 中只是将不同分辨率的特征对齐后堆叠, 忽略了低层特征包含较多局部位置信息, 高层特征包含更多全局语义信息, 而忽略这些特征之间的联系将不可避免影响检测的性能。因此考虑不同分辨率特征之间的相关性, 自适应对通道相关性建模, 通过在全局语义信息融合局部位置信息来提高检测性能是一个重要问题。同时, 从人类视觉出发, 对于不同尺度的目标需要不同大小感受野的特征去识别, 神经网络的高层特征中包含更丰富的语义信息, YOLOv3-SPP[7](MobileNets based on depthwise separable convolutions) 算法中通过对高层语义特征增强感受野, 可以加强网络的特征提取能力, 虽然其中 SPP(spatial pyramid pooling) 网络能够捕获上下文信息, 但同时破坏了图像中的姿态和空间等信息, 造成部分高维特征丢失的问题。

针对目标检测中的多尺度问题, 本文提出一种自适应上下文特征的多尺度目标检测算法。首先, 基于改进的注意力机制设计了特征融合网络 A-PANet(attention-path aggregation network), 自适应地调整通道间的相关性和不同分辨率特征的通道权值, 实现局部特征和全局特征的融合, 提升检测的精度。其次, 设计了多尺度感受野特征金字塔网络 MSPNet(multi sensory pyramid network), 利用不同膨胀率的卷积, 从高层语义特征中学习到不同大小感受野的特征, 识别不同尺度的物体, 提高检测的精度。通过对 PASCAL VOC[8] 数据集上的实验结果进行分析评估, 本文的方法相较于其他先进算法的性能有了显著提高。

1　算法原理

1.1　网络结构设计

本文算法从多尺度感受野和自适应特征融合两方面, 设计了自适应上下文特征的多尺度目标检测算法。方法的整体框架如图 1 所示。具体来说, 以 Darknet53 作为主干特征提取网络, 首先, 将图像输入主干特征提取网络, 获取高层特征 P_0、次高层特征 P_1、浅层特征 P_2; 其次, 为了从高层语义特征中挖掘标签中的上下文信息, 设计了 MSPNet 网络, 从高层特征 P_0 中提取多尺度感受野特征, 并通过 3 次卷积实现上下文信息的融合; 最后, 基于改进的注意力机制 SE*, 设计了 A-PANet 网络, 对不同分辨率特征 P_0、P_1、P_2 进行加权融合, 实现局部特征和全局特征的融合, 并利用融合后的多尺度特征对不同尺度物体实现分类和回归。

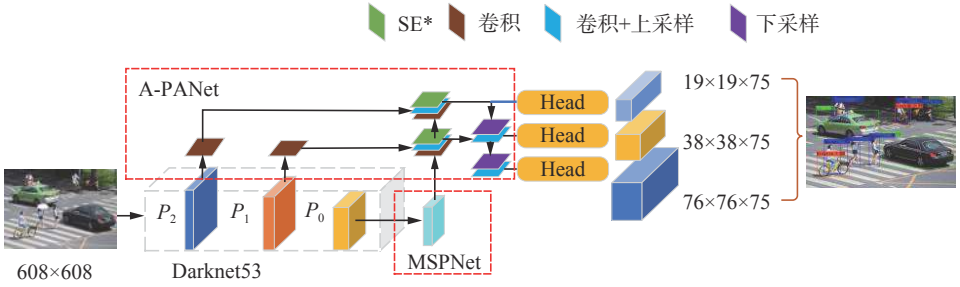

图 1　算法框架结构

Fig. 1　Algorithm framework structure

1.2　多感受野特征提取网络

从人类视觉出发, 识别物体的类别, 除了当前物体的外观特征, 还需要周围环境作为辅助 (比如汽车和人通常同时出现, 椅子在桌子附近)。如

何从高级语义特征中获取不同尺度目标的语义信息，并获取上下文信息辅助识别小目标物体，是提高检测性能的关键问题。针对此问题，本文提出多感受野的特征提取网络 MSPNet，利用多分支并行空洞卷积，通过不同大小的感受野，从高层特征 P_0 中挖掘不同尺度目标的语义信息，并通过融合不同尺度的感受野特征从标签中获取上下文信息。

MSPNet 的网络结构如图 2 所示。首先，以主干特征提取网络输出的高层语义特征 $x \in \mathbf{R}^{W\times H\times C}$ 作为输入，其中 W、H 为特征图的宽高，C 为特征维度。其次，将高层语义特征 x 分别经过 3 个膨胀系数为 τ 的多感受野特征提取分支，$x^i_{i=1,2,3} \in \mathbf{R}^{W'\times H'\times C'}$ 表示每个分支捕获的不同大小感受野以及不同尺度的特征信息。其中 W'、H' 和输入特征图的宽高 W、H 保持一致，C' 下降为输入通道的 1/16。然后，对 3 个分支进行归一化处理，加快网络的训练以及收敛速度防止梯度爆炸，并使用 Leaky_ReLU 激活函数增加非线性。最后，使用 1×1 的卷积核对高层语义特征 x 进行卷积处理，输出特征 $x^4 \in \mathbf{R}^{W'\times H'\times C'}$ 并作为残差结构和其他分支获得的特征进行感受野从小到大的堆叠，输出多感受野增强提取特征，再使用 3 次卷积进行多感受野特征加强融合。得到最终加强多感受野特征 $x' \in \mathbf{R}^{W\times H\times C}$。

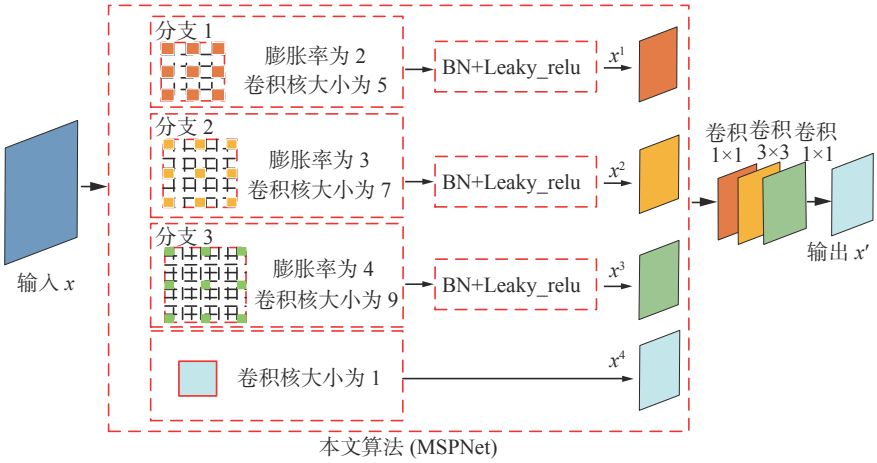

图 2　多感受野特征提取网络 (MSPNet) 结构

Fig. 2　Stucture of multi-receptive field feature extraction network (MSPNet)

不同尺度的目标需要不同大小感受野的特征识别，YOLOv3-SPP[7] 算法中 SPP 网络利用多分支池化提取不同大小感受野特征，而池化会造成语义特征的丢失问题。为了获取多尺度的感受野且不造成特征的丢失，本文提出利用膨胀卷积增加感受野，保持特征的尺寸不变化。膨胀卷积[9-10]通过稀疏采样的方式进行卷积，通过在卷积核内部填充权值为 0 的参数使得卷积核的感受野增大且不会增加额外参数。膨胀卷积后有效感受野大小如式 (1) 所示。

$$k' = k + (\tau - 1)\times(k - 1) \tag{1}$$

式中：k 代表原卷积核大小；τ 代表膨胀率；k' 表示有效感受野。神经元的感受野越大表示和原始图像的接触范围越大，提取的信息则是更加全局，包含语义层次更高的特征；感受野越小则是提取的特征趋向于局部和细节。本文设计 3 个膨胀卷积分支，以膨胀率分别为 2、3、4 的 3×3 卷积核，根据式 (1)，每个分支对高层语义特征的有效感受野大小分别是 5×5、7×7、9×9。输出 x' 为以金字塔形式堆叠的多感受野特征，其表达式为

$$x' = f[F(W^1x \oplus W^2x \oplus W^3x + W^4x)] \tag{2}$$

式中：W^1、W^2、W^3 为每个分支的学习参数；$\oplus$ 为特征的堆叠；F 为三次卷积；x 为输入高语义特征；W^4 为残差边的学习参数；f 为非线性激活函数 Leaky_ReLu，其表达式为

$$f(x) = a^*x \tag{3}$$

其中 a 为超参数，通常取值为 0.01，在反向传播过程中，对于 Leaky_ReLU 激活函数输入小于零的部分，也可以计算得到梯度，避免梯度消失的问题。

1.3　自适应特征融合网络 A-PANet

神经网络的低层特征包含丰富的目标位置信息，高层特征则包含目标的语义信息。考虑不同尺度目标的语义特征出现在不同分辨率特征图中，在全局语义特征中有效地融合局部位置特征，是解决检测中多尺度问题的关键。本文提出一种自适应特征融合网络 A-PANet，通过自适应地调整不同分辨率特征间的依赖性，实现语义特征和位置特征的有效融合。

A-PANet 网络结构如图 3 所示。其中，C_4、C_5、C_6、$C_7 \in \mathbf{R}^{W\times H\times C}$ 为主干特征提取网络输出的不同分辨率特征，其中 W、H、C 分别为每个特征的宽度、长度和通道数。⊕ 表示不同分辨率特征自适应融合模块，其具体实现如图 4。P_5、P_6、$P_7 \in \mathbf{R}^{W'\times H'\times C'}$ 为不同分辨率特征自适应融合后的特征，其尺度分别为 76×76、38×38、19×19，每个网格点输出 3 个预测框，分别用于检测不同尺度大小的物体。它每一层预测所用的特征图都融合了不同分辨率、不同语义强度的特征，融合的不同分辨率特征图对应不同大小的物体检测。

通道注意力[11-12]在于分配各个卷积通道之间的资源，可以理解为让神经网络明白在看什么，网络可以有选择性地加强包含重要信息的特征并抑制作用无关或较弱关联的特征。图 3 中 ⊕ 为特征自适应融合模块，本文针对通道注意力机制实现以下改进。首先，为了提高通道间的非线性拟合能力，对通道注意力 SE[11] 的 FC 层进行不降维处理，避免降维造成的细节信息损失。其次，针对采用两个不降维 FC 会显著增加模型复杂度的问题，本文利用两个不降维的一维卷积代替 FC 层，降低模型复杂度的同时保持性能增益，具体实现如图 4 虚线框所示。

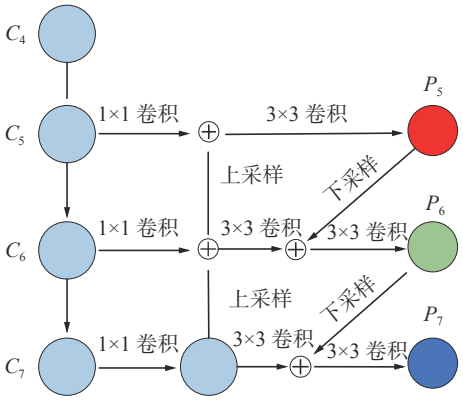

图 3　自适应特征融合网络 A-PANet 结构

Fig. 3　Structure of adaptive feature fusion network A-PANet

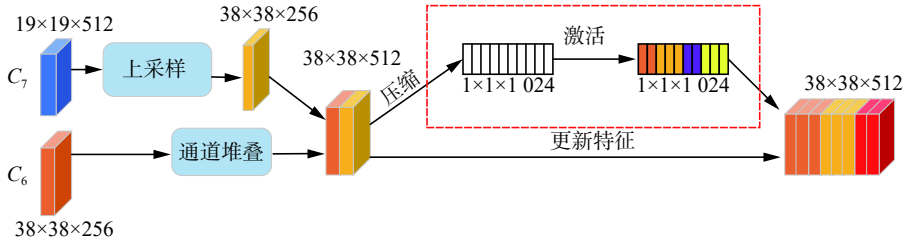

图 4　特征自适应融合模块结构

Fig. 4　Structure of feature adaptive fusion module

图 4 以 C_7 和 C_6 两个不同的分辨率特征融合为例说明，其余融合方式和此一致。首先，对 C_7 特征进行双线插值上采样，恢复其宽高并和 C_6 层特征图的宽高保持一致。其次，对输入进行压缩，利用全局平均池化，将输入的二维特征图变成单个像素值且通道数不发生变化，输出的每个特征通道上具有全局的感受野；最后，通过两次不降维的一维卷积，并在激活函数前引入 BN 层加速收敛，增加网络通道间的非线性拟合能力。

特征自适应融合模块如式 (4)：

$$\omega = \sigma(f\{w_1,w_2\}(g(x))) \tag{4}$$

式中：σ 为 sigmoid 激活函数；x 为输入特征图。$g(x)$ 为全局池化函数如式 (5) 所示，其功能是对输入特征图的每个通道进行全局平均池化 (GAP)。其中 W、H 分别表示输入特征图的宽高。

$$g(x) = \frac{1}{W\cdot H}\sum_{i=1,j=1}^{W,H} x_{ij} \tag{5}$$

$f\{w_1,w_2\}$ 函数的作用如式 (6)：

$$f\{w_1,w_2\}(g(x)) = w_2\mathrm{ReLU}(w_1 * g(x)) \tag{6}$$

式中：w_1 表示第一个卷积层的可学习参数；w_2 为经过第 2 个卷积层的可学习参数；*表示为逐元素相乘。这个模块负责构建通道的相关性以及自适应地为不同通道学习到不同的通道注意力权重。通过对特征通道间的相关性进行建模，网络专注于更有用的通道并增强辨别学习能力。

2　实验结果与分析

2.1　数据集和实验环境

本实验所用的数据集为图像识别和分类的标准化数据集 PASCAL VOC，数据集标签中包含 20 个类别，它是常用于目标检测任务的训练和评价的公开数据集。此数据集包含行人、车辆、生活物品等 20 个语义类别高精度标注图像。本实验使用的训练集是包含 VOC2007 的训练和验证集及 VOC2012 的训练和验证集的联合训练集一共有 16 551 张图，取其中 90% 为训练集，10% 作为验证集。测试集选取 VOC2007 的测试集，一共 4952 张图。本实验的环境配置如表 1。

表 1　实验环境配置

Table 1　Experimental environment configuration

实验配置	型号或参数
运行环境	Windows 10
CPU	I9-10900k
显卡	GeForce RTX 2080ti
显卡内存/GB	11
CUDA版本	10.0
深度学习框架	Pytorch1.2
算法语言	Python

2.2　实验评价标准

在目标检测任务中，检测精度 (average precision, AP) 体现每个物体种类的检测精度，mAP (mean average precision) 是对所有检测种类的 AP 进行算数平均，用来衡量整个网络的检测精度，mAP 值越大则检测检测精度越高。AP 是由检测精度 (precision) 和召回率 (recall) 组成的 PR 曲线面积计算得出。精度 (P) 和召回率 (R) 的计算方法为

$$P=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \tag{7}$$

$$R=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{8}$$

式中：TP 表示为正样本且预测结果为正样本的检测框；FP 表示为负样本但预测结果为正样本的检测框；FN 为负样本且检测结果为负样本的检测框。

2.3　实验参数设置

为了验证改进后的网络模型的检测精度变化，在相同的实验环境以及使用的数据集都为 VOC2007+VOC2012 的联合训练，并在 VOC2007 的测试集上计算每个类的 AP 值以及 20 个类的 mAP。整个训练过程中使用了迁移学习的思想，利用在大型数据集上预训练的模型参数对网络进行参数初始化，来加速推理和提高网络收敛速度。实验具体参数设置如表 2 所示。

表 2　实验参数设置

Table 2　Experimental parameter setting

迭代次数	批次	学习率	权重衰减
0~25	4	0.001	5×10^{-4}
25~50	2	0.000 1	5×10^{-4}

训练一共设置 50 个 epoch，训练的前 25 个 epoch 对网络的主干特征提取网络的部分参数进行冻结训练，后 25 个 epoch 解冻后整体训练。网络采用批量随机梯度下降法来优化损失函数，前 25 个 epoch 设置 Batch_size 为 4，初始学习率为 0.001，权重衰减率为 0.000 5，解冻训练后继续训练 25 个 epoch，此时网络学习率设定为 0.000 1，Batch_size 为 2，权重衰减为 0.000 5。通过测试 loss 的变化情况可以看到网络模型的拟合情况，并选取达到最佳拟合效果的 epoch 作为网络的权重。

2.4　PASCL VOC 上的定量评价

Faster-RCNN[1-2]、YOLO[13-14]、SSD[15]、DSSD321[16] 等都是目标检测领域常用的几种检测算法，本文将几种算法都在 VOC2007+VOC2012的混合数据集上训练以及使用 VOC2007 测试集为测试数据，其中分别列举了 Faster-RCNN、SSD、R-FCN[17] 的实验对比结果。其中 SSD321、SSD300 除了输入图片的大小不同其他设置都完全相同，基线模型 YOLOv3+来自文献 [14]。表 3为不同算法在 VOC2007 上得到的测试结果。

表 3　各种算法在 VOC2007 上的测试效果

Table 3　Test effects of various algorithms on VOC2007

算法分类	算法	主干网络	输入大小/Pixel	mAP/%
双阶段目标检测算法	Faster-RCNN[1]	VGG-16	~1 000×600	73.2
	Faster-RCNN[2]	Residual-101	~1 000×600	76.4
	R-FCN[17]	Residual-101	~1 000×600	80.5
单阶段目标检测算法	SSD300[15]	VGG-16	300×300	74.3
	DSSD321[16]	Residual-101	321×321	78.6
	YOLOv2[13]	DarkNet53	416×416	76.8
	YOLOv3[14]	DarkNet53	416×416	80.25
	YOLOv3+[14]	DarkNet53	416×416	83.68
	本文算法	Darknet53	416×416	84.94
	本文算法	Darknet53	608×608	**85.74**

从表 3 不同算法测试得到的 mAP 的数据对比发现, 本文提出的基于自注意力和多尺度特征融合的目标检测方法在检测精度上具有更好的表现。在 VOC2007 数据集上的检测精度, 相较于双阶段目标检测算法, 如以 VGG-16 和 Residual-101 作为主干提取网络的 Faster-RCNN, 检测精度分别提升了 12.54% 和 9.34%, 相较于单阶段的目标检测算法 DSSD321 和 YOLOv3, 检测精度分别提升 7.14% 和 5.49%。相较于本文的基线模型 YOLOv3+仍有 2.06% 的提升,

为了验证本文算法在解决目标检测中多尺度问题上的优越性, 将本文算法和其他多尺度目标检测算法的检测结果进行比较分析, 实验结果如表 4 所示。

表 4　多尺度目标检测算法在 VOC2007 数据集上的测试结果

Table 4　Test results of multi-scale target detection algorithm on VOC2007 dataset

算法	主干网络	训练数据	mAP/%
SSD[18]	VGG	VOC2007+2012	76.7
RefineDet512+[19]	VGG	VOC2007+2012	83.8
RFBNet512[20]	VGG	VOC2007+2012	82.2
本文算法	Darknet	VOC2007+2012	**85.7**

表 4 中 SSD 算法是以主干提取网络的不同特征层检测不同尺度物体, 解决多尺度目标检测问题; RefineDet512+算法是基于 SSD 和 FPN 算法的改进, 通过不同特征层间的融合检测出不同尺度物体。RFBNet 算法从感受野的角度出发, 在 SSD 算法基础上对不同特征层使用 RFB 网络增加感受野, 提升多尺度物体的检测能力。本文算法针对多尺度问题同时从感受野和多尺度特征融合的角度出发, 首先, 提出多分支的并行空洞卷积网络 MSPNet 对不同感受野信息融合, 挖掘上下文信息。其次, 基于注意力机制提出自适应特征融合网络 A-PANet, 考虑不同特征层间的相关性实现多尺度特征融合。实验结果证明, 本文算法相较于其他多尺度目标检测算法性能上得到显著提升。

检测速度也是衡量检测算法性能的重要指标之一, 本文对比不同检测算法在 VOC2007 数据集上的测试速度 FPS。为了公平比较, 在测试阶段, 设置批次大小为 1, 各算法的时间性能对比见表 5。

表 5　不同算法在 VOC2007 数据集上的测试速度

Table 5　Different algorithms test speeds on the VOC2007 dataset

算法	网络	速度/(f·s^{-1})	显卡
Faster-RCNN[1]	VGG-16	7.0	TitanX
Faster-RCNN[2]	Residual-101	2.4	K40
SSD300[15]	VGG-16	46.0	TitanX
DSSD321[16]	Residual-101	9.5	TitanX
YOLOv3+[14]	DarkNet53	27.83	2080Ti
本文算法(416)	DarkNet53	27.47	2080Ti
本文算法(608)	DarkNet53	23.08	2080Ti

考虑到平台差异对检测速度的影响, 本文在此实验平台上对基线模型进行复现, 算法速度达到 27.83 f/s。本文算法由于增加额外的计算, 相较于基线模型, 当输入图片大小为 416×416, 检测时间多消耗约 0.3 s, 当输入图片大小为 608×608, 多消耗 17% 的时间。但由表 5 可以看出, 本文算法的检测速度明显高于双阶段目标检测算法, 同时, 由于硬件平台的限制, 检测速度略低于其他单阶段目标检测算法。如图 5, 综合来看, 本文算法综合效率最高, 既实现了更高的检测精度, 又保持速度上的优势。

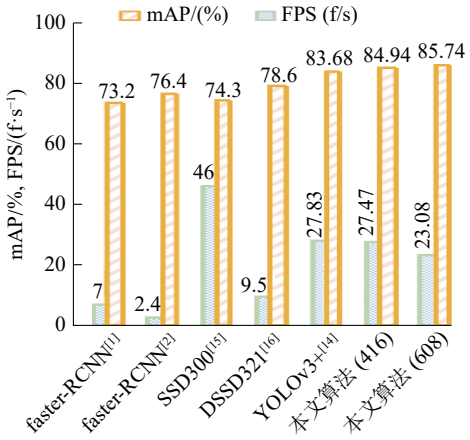

图 5　VOC 2007 测试集上的 time-mAP 对比

Fig. 5　Time-mAP comparisons on VOC 2007 test set

2.5　消融实验

为了验证本文算法具有更好的检测精度, 分别评估了多感受野特征提取网络和自适应特征融合网络, 并研究了多感受野特征提取网络的分支数量对实验影响, 设定在相同的实验环境和 VOC 数据集进行消融实验, 并对实验结果进行分析。具体实验分为 3 部分:

1) 基线模型中单独验证自适应特征融合网络;

2) 基线模型中单独验证多感受野特征提取网络;

3) 基线模型中同时引入自适应特征融合网络和多感受野特征提取网络。

为了验证本文算法的优越性, 实验以在 VOC 2007 的测试集上检测结果为基准, 独立验证每个模块对模型的检测精度的影响, 统计结果如表 6 所示。由于文献 [14] 中未公布每个类的 AP, 为了公平比较, 本文复现了每个类的 AP, 并且 mAP 和文献 [14] 无差距。此外, 为了验证多感受野网络的有效性, 分别对多感受野特征提取网络 MSPNet-T、MSPNet-F(其分支数量为 3 和 4) 进行实验。

表 6　在 VOC2007 数据集上的消融实验结果

Table 6　Detection effect of improved algorithm on VOC2007　　%

方法	YOLOv3+	本文算法			
		A-PANet	MSPNet(F)	MSPNet(T)	A-PANet+MSPNet(T)
mAP	83.66	85.67	84.90	85.36	**85.74**
Aero	89.07	92.72	91.65	**93.54**	92.97
Bike	89.13	92.46	92.96	92.68	**93.31**
Bird	85.24	85.48	83.86	85.03	**86.21**
Boat	73.43	**79.68**	76.28	78.97	75.26
Bottle	79.80	79.46	81.33	80.92	**81.73**
Bus	87.83	**92.42**	91.80	91.98	92.02
Car	93.77	94.98	94.67	94.85	**95.50**
Cat	89.84	**89.93**	88.61	88.57	88.53
Chair	70.19	71.01	72.59	71.59	**74.06**
Cow	89.11	**93.07**	92.69	91.46	93.03
Table	78.69	**79.23**	76.65	76.92	78.39
Dog	87.40	87.54	85.02	**88.70**	88.40
Horse	89.29	91.18	90.71	**91.29**	90.69
Motorbike	89.75	**92.52**	91.98	91.00	91.44
Person	89.38	91.24	91.00	90.98	**91.84**
Plante	59.43	58.59	59.24	59.21	**60.43**
Sheep	79.74	**89.51**	88.50	87.37	87.64
Sofa	77.20	75.46	75.80	74.49	**77.40**
Train	90.87	89.97	87.13	**91.80**	89.05
Tv	84.00	86.81	85.57	85.80	**86.83**

从表 6 中的每个类的 AP 数据得到以下结果: 特征融合网络在全局语义特征中融合了局部位置特征, 相较于基线模型, 在大目标物体上的性能有显著提升, 如 Boat(船)、Cow(牛)、Sheep(羊) 等, 尤其是在羊和船的 AP 值提升接近 10%。得到这样的实验结果是合理的, 因为大目标物体通常在图片中占有较多像素, 因此在卷积神经网络的高层特征中包含丰富的语义信息, 识别精度高。但同时因为物体较大, 在卷积神经网络下采样时造成位置信息的偏移, 影响回归精度。本文提出的特征融合网络, 通过注意力模块可以自适应地调整各通道的特征响应值, 通过学习参数的方式来自动获取到每个特征通道的重要程度, 然后依照这个重要程度去改变不同分辨率特征的通道权重。有效地为大目标物体在语义信息中融合了位置信息, 进一步提升大目标物体的检测能力, 展示了本文特征融合网络的优越性。

多感受野特征提取网络从高层语义特征中挖掘标签中相互依赖的全局语义信息。多感受野特征提取网络的性能相较于基线模型提高了 1.70%, 同时对于 bike(自行车) 和 person(人) 的 AP 有 3.5% 和 1.5% 的提升。因为客观世界中这类目标通常具有很强的依赖关系, 它们往往同时出现, 进一步说明, 本文提出的多感受野特征提取网络能有效地聚集上下文信息, 提升检测精度。

从最终的实验结果中可以看出, 相较于基线模型, 本文算法有效地提升了不同尺度物体的检测精度, 如小目标物体 cat(猫) 和 dog(狗) 的 AP 值有 1% 左右的提升, 大目标物体 bike(自行车) 和 bus(公交车) 有 4% 左右的提升。在 VOC2007 数据集上的可视化测试结果如图 6 所示, 从定量实验结果来看, 本文算法在处理不同尺度物体问题

上的合理性和有效性得到充分证明。图 6 中包括每一个类的 AP 以及 20 个类的 mAP, 其中图 6(a) 为基线模型的 mAP, 图 6(b) 为本文算法的 mAP。从 mAP 的对比可以发现改进后的网络在多个种类的物体的检测精度上都相较于原始网络具有不错的提升效果。

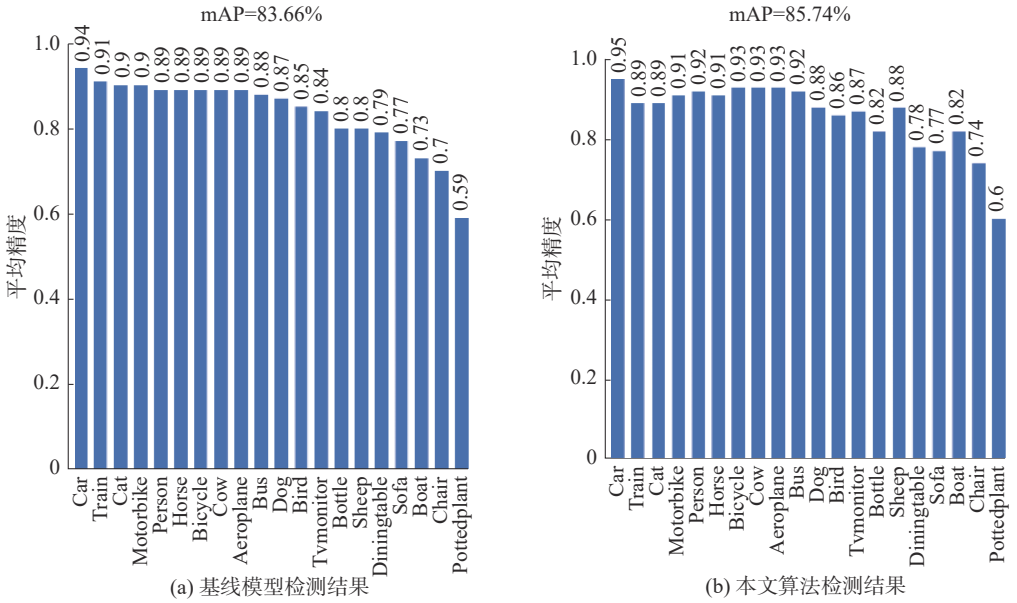

图 6　在 VOC 2007 数据集上的测试结果

Fig. 6　Test results on VOC 2007 datasets

2.6　定性评价结果

为了更加直观地评价本文算法, 图 7 给出了在 VOC2007 的测试集上的实验结果对比。其中 1、3 列为基线模型的测试结果, 第 2、4 列为本文算法的测试结果。

图 7　基线模型和本文算法在 VOC 数据集上实验结果

Fig. 7　Experimental results of baseline model and algorithm in this paper on VOC dataset

对图 7 的可视化检测结果中进行定性分析：图 7(a)、(d) 为从 VOC2007 数据集中随机选取室内图片，原算法存在对椅子的漏检问题，而本文算法利用多感受野特征提取网络，聚集上下文信息，通过挖掘标签中的关系 (如桌椅通常同时出现)，减少物体的漏检；针对一幅图片中需要检测不同尺度物体的问题，如图 7(b) 中杂志上的人和图 7(g) 中远处的马，原算法在小目标物体上存在漏检以及误检的问题，本文算法通过自适应特征融合网络，在语义信息中融合位置信息，有效改善了不同尺度物体检测问题；虽然本文算法比原始算法检测精度更高，但仍存在当目标和背景特征相似时 (图 7(f) 中的黄羊)，网络无法识别出物体的问题。综合来看，本文算法的检测性能更优异。

3　结束语

本文提出一种结合上下文特征和自适应特征融合的目标检测算法。首先利用主干特征提取网络 Darknet53 获取不同尺度的特征图，接着构建一种多感受野特征提取网络，从高层语义特征中聚集上下文特征，挖掘标签中隐含的全局知识。最后构建一种自适应特征融合网络，结合通道注意力机制，实现不同分辨率特征的融合，在不同尺度的特征图中融合全局语义信息和局部位置信息。在 PASCAL VOC 数据集上的实验结果表明，本文算法既能保持速度的优势，同时有效地提升了不同尺度物体的检测精度，更具有实用价值。在下一步的工作中，将继续改进模型，探索解决物体和背景特征相似不易识别的问题。

参考文献：

[1] 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.

[2] 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. Las Vegas, USA, 2016: 770−778.

[3] SINGH B, DAVIS L S. An analysis of scale invariance in object detection-SNIP[C]//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 3578−3587.

[4] SINGH B, NAJIBI M, DAVIS L S. SNIPER: efficient multi-scale training[C]//Proceedings of the 32nd Conference on Neural Information Processing Systems. Montréal, Canada, 2018: 9333−9343.

[5] LIN T Y, DOLLÁR P, GIRSHICK R, et al. Feature pyramid networks for object detection[C]//Proceedings of the 30th IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2017: 936−944.

[6] REDMON J, FARHADI A. YOLOv3: an incremental improvement[EB/OL]. (2018-04-08)[2021-01-01].http://arxiv.org/abs/1804.02767.

[7] HOWARD A G, ZHU M, CHEN B, et al. MobileNets: efficient convolutional neural networks for mobile vision applications[EB/OL]. (2017-04-14)[2021-01-01].http://arxiv.org/abs/1704.04861.

[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.

[9] YU F, KOLTUN V. Multi-scale context aggregation by dilated convolutions[C]//Proceedings of the 4th International Conference on Learning Representations. San Juan, Puerto Rico, 2016.

[10] LI Yanghao, CHEN Yuntao, WANG Naiyan, et al. Scale-aware trident networks for object detection [C]//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South), 2019: 6053−6062.

[11] HU Jie, SHEN Li, ALBANIE S, et al. Squeeze-and-excitation networks[J]. IEEE transactions on pattern analysis and machine intelligence, 2020, 42(8): 2011–2023.

[12] WANG Qilong, WU Banggu, ZHU Pengfei, et al. ECA-Net: efficient channel attention for deep convolutional neural networks[C]//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 11531−11539.

[13] 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. Las Vegas, USA, 2016: 779−788.

[14] ZHANG Zhi, HE Tong, ZHANG Hang, et al. Bag of freebies for training object detection neural networks [EB/OL]. (2019-04-12)[2021-01-01]. http://arxiv.org/abs/1902.04103.

[15] 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.

［16］ FU Chengyang, LIU Wei, RANGA A, et al. DSSD: deconvolutional single shot detector[EB/OL]. (2017-01-23)[2021-01-01].http://arxiv.org/abs/1701.06659.

［17］ 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.

［18］ YUN S, HAN D, CHUN S, et al. CutMix: regularization strategy to train strong classifiers with localizable features[C]//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South), 2019: 6022−6031.

［19］ ZHANG Shifeng, WEN Longyin, BIAN Xiao, et al. Single-shot refinement neural network for object detection[C]//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 4203−4212.

［20］ LIU Songtao, HUANG Di, WANG Yunhong. Receptive field block net for accurate and fast object detection[C]// Proceedings of the 15th European Conference on Computer Vision. Munich, Germany, 2018: 404−419.

作者简介：

王凤随，副教授，主要研究方向为视频通信、计算机视觉。承担国家自然科学基金、安徽省自然科学基金等多项课题研究。发表学术论文 40 余篇。

陈金刚，硕士研究生，主要研究方向为图像目标检测与识别。

王启胜，硕士研究生，主要研究方向为图像目标检测与识别。

中国人工智能学会关于征集 2022 重大科学问题、工程技术难题和产业技术问题的通知

学会各分支机构、地方人工智能学会、会员单位、会员、人工智能领域科技工作者：

为进一步加强科技前瞻研判，引领原创性科研攻关，推进科技自立自强，根据《中国科协办公厅关于征集 2022 重大科学问题、工程技术难题和产业技术问题的通知》（科协办函创字〔2022〕19 号），中国人工智能学会现开展人工智能领域前沿科学问题、工程技术难题和产业技术问题征集工作。现就有关事项通知如下：

一、征集时间

从通知发布之日起，至 2022 年 3 月 31 日止。

二、征集内容和领域

面向世界科技前沿、面向经济主战场、面向国家重大需求、面向人民生命健康，征集对未来科技发展具有引领作用的前沿科学问题、工程技术难题和产业技术问题。加强有关国家战略科技力量和战略性新兴产业的科技问题征集，尤其是重大基础研究问题、关键共性技术、前沿引领技术、现代工程技术、颠覆性技术、“卡脖子”技术、促进可持续发展关键技术等问题，重点关注前沿交叉融合领域的相关问题难题。征集范围原则上覆盖所有自然科学、工程技术与产业领域，重点征集信息科技领域。

三、征集方式

推荐单位可单独或者联合推荐，鼓励联合相应国外科技组织或国际专家共同推荐，每个推荐单位可推荐前沿科学问题、工程技术难题和产业技术问题各 3~5 个。每位会员或人工智能领域科技工作者可推荐前沿科学问题、工程技术难题和产业技术问题各 1~3 个。

中国人工智能学会可向中国科协推荐前沿科学问题、工程技术难题和产业技术问题各 3~5 个。

四、联系方式

联系人：贾老师

电　话：010-82686686

邮　箱：zhb@caai.cn