控制与决策
Control and Decision

自适应感受野网络的行人重识别

王松，纪鹏，张云洲，朱尚栋，暴吉宁

引用本文:
王松, 纪鹏, 张云洲, 等. 自适应感受野网络的行人重识别[J]. 控制与决策, 2022, 37(1): 119−126.

在线阅读 View online: https://doi.org/10.13195/j.kzyjc.2020.0505

自适应感受野网络的行人重识别

王 松[1], 纪 鹏[1†], 张云洲[1,2], 朱尚栋[1], 暴吉宁[2]

(1. 东北大学 机器人科学与工程学院，沈阳 110169；2. 东北大学 信息科学与工程学院，沈阳 110004)

摘 要: 行人重识别通常删除特征提取网络中的最后一个空间下采样操作,以增加最后输出特征图的分辨率,保留更多的细粒度特征. 然而,这种操作会大幅减小神经网络的感受野,而更大的感受野可以为行人重识别提供更多的上下文信息. 同时,在实际的视觉皮层中,相同区域的神经元的感受野是不同的,但当前行人重识别网络的设计大多忽视了这一点. 为了解决上述问题,提出一种新颖的自适应感受野网络. 网络的设计受启发于生物的视觉系统,通过在多分支网络上设置不同大小的感受野,结合注意力机制让网络自行选择合适的感受野特征,从而实现网络感受野的自适应,并且采用分组卷积使得自适应感受野模块更加轻量级. 同时在各个分支利用空洞卷积增大感受野,补偿删除最后下采样操作所减少的网络感受野. 在公开的大规模数据集上进行实验,实验结果表明,所提出的算法相比于基线方法有显著的提升,当使用ResNet-50作为特征提取网络时,在DukeMTMC-reID、Market-1501数据集上的Rank-1和mAP分别达到89.2 %和76.0 %、95.2 %和87.2 %. 与现有方法相比,所提出算法在精度上有明显的提升.

关键词: 行人重识别；深度学习；自适应感受野；注意力机制；空洞卷积；分组卷积

中图分类号: TP242 **文献标志码:** A

DOI: 10.13195/j.kzyjc.2020.0505 **开放科学(资源服务)标识码(OSID):**

引用格式: 王松,纪鹏,张云洲,等. 自适应感受野网络的行人重识别[J]. 控制与决策, 2022, 37(1): 119-126.

Adaptive receptive network for person re-identification

WANG Song[1], *JI Peng*[1†], *ZHANG Yun-zhou*[1,2], *ZHU Shang-dong*[1], *BAO Ji-ning*[2]

(1. Faculty of Robot Science and Engineering，Northeastern University，Shenyang 110169，China；2. College of Information Science and Engineering，Northeastern University，Shenyang 110004，China)

Abstract: Person re-identification typically removes the last spatial down-sampling operation in the backbone to increase the resolution of the final output feature map and preserve more fine-grained features. However, this operation substantially reduces the size of a receptive field, and a larger receptive field can provide more contextual information for person re-identification. At the same time, in the actual visual cortex, the receptive field of neurons in the same region are different, but this is largely ignored by the current design of pedestrian recognition networks. To solve the above problems, this article proposes a novel adaptive receptive field network. The design of the network is inspired by the visual system of living organisms. By setting a different sized receptive field on the multi-branch network, combined with the attention mechanism to allow the network to select the appropriate receptive field characteristics, the network receptive fields adaptive is realized, and the use of packet convolution makes the adaptive receptive field module more lightweight. The receptive field is also increased in each branch using empty convolution to compensate for the reduction of the network receptive field by deleting the last downsampling operation. Experiments are performed on publicly available large-scale datasets, and results show that the algorithm has a significant improvement over the baseline approach, with Rank-1 and mAP on the DukeMTMC-reID, Market-1501 datasets reaching 89.2 % and 76.0 %, 95.2 % and 87.2 %, respectively, when using ResNet-50 as backbone. Compared with the existing methods, the proposed algorithm also has a significant improvement in accuracy.

Keywords: person re-identification；deep learning；adaptive receptive field；attention mechanism；dilated convolutions；grouped convolutions

收稿日期: 2020-05-01；**修回日期:** 2020-10-05.

基金项目: 中央高校基本科研业务费专项资金项目(N172608005, N182608004)；国家自然科学基金项目(61973066, 61471110)；装备预研领域基金项目(61403120111)；航天系统仿真重点实验室基金项目(6142002301).

责任编委: 薛建儒.

†通讯作者. E-mail: jipeng@mail.neu.edu.cn.

0 引 言

行人重识别[1]也称行人再识别,是利用计算机视觉技术判断图像和视频序列中是否存在特定的行人,被广泛认为是图像检索的子问题. 由于监控摄像头所获得的人脸部信息通常较为模糊,难以利用人脸识别的方法进行行人的识别搜索[2]. 在实际环境中,行人所在的环境与摄像头摆放的位置不同造成图像抓拍的效果存在很大的差异,导致行人的姿势、轮廓和颜色等特征差别很大. 尤其是当物体遮挡行人时,行人重识别的难度将大大增加.

随着智能城市建设的快速发展,行人重识别领域引起了极大关注. 文献[3-4]最早在行人重识别领域引入深度学习方法. 按照网络输出的特征类型,行人重识别可以分为基于全局特征和局部特征方法. 全局特征一般是对特征图进行全局池化直接得到,推理速度更快,更为简单实用. 然而,为了追求在数据集上的表现,当前的行人重识别模型通常采用局部特征或者局部特征结合全局特征进行检索,仅利用全局特征进行检索的模型较少. 为了提高行人重识别的性能,目前的方法[5]往往设计复杂的网络结构并连接多个支路的局部特征,导致运行速度很慢,难以投入实际应用. 此外,基于局部特征的方法常需要进行人体对齐的操作,因为只有将相对应的身体部位进行比较才会有好的效果. 相对地,利用全局特征的方法往往非常高效简洁,并没有很多的定制化设计,也不需要身体对齐这一条件. 因此,在行人不对齐情况下,整体特征相比于利用区域划分的局部特征更加鲁棒,适合于实际的行人重识别应用.

行人重识别是在跨摄像头的条件下进行行人图像细粒度检索的任务,丰富的细粒度特征是非常关键的. 更高分辨率的特征图可以带来更为丰富的细粒度特征,显著提升行人重识别模型的精度. 文献[5]提出了删除特征提取网络中的最后一个空间下采样操作,以增加最后输出特征图的分辨率,保留更多细粒度特征. 这种操作仅仅增加非常少量的计算成本,却有着非常显著的效果,但同时也存在很大的问题:取消最后的下采样操作会使得网络的感受野大大减小,而对于分类任务,网络感受野通常是越大越好. 大的感受野可以使得网络获取更多的上下文信息,这对于行人重识别至关重要.

神经网络的感受野在各种计算机视觉任务上得到了广泛的应用. 初级视觉皮层(V_1)神经元[6]的局部感受野(RFs)激发了卷积神经网络(CNNs)的出现[7],并持续推动着现代CNN结构的改变. 例如,在视觉皮层中,相同区域(例如V_1区域)中的神经元的感受野区域尺寸是不同的,这使得神经元能够在处理阶段收集多尺度空间信息. 在卷积神经网络(CNN)中,该机制已被广泛采用,较具代表性的实例是InceptionNets[8],它融合了来自不同分支中的多尺度信息. 本文提出的自适应感受野模块是通过视觉注意力机制[9]动态选择合适的感受野特征,有效融合了不同分支的多尺度信息,并利用空洞卷积[10]进一步增大感受野. 自适应感受野模块是轻量级的,计算成本和模型参数仅有少量的增加,但效果提升非常明显.

针对行人重识别模型易受遮挡和尺度变化影响的问题,本文提出一种简洁高效的自适应感受野网络,进一步加强网络的特征提取能力. 本文的主要贡献包括:

1)考虑到行人重识别网络往往采用删除特征提取网络中的最后一个空间下采样操作以保留更多细粒度特征,但是这一操作会减少网络的感受野,使得网络难以捕捉更有效的上下文信息. 本文采用空洞卷积操作,在不增加参数量和计算量的情况下,弥补感受野的不足.

2)神经元的感受野是动态变化的,从而能够关注不同大小区域的信息,例如为了判定不同的行人,有时关注整体的外观,有时聚焦于某一特定区域进行比较. 目前的行人重识别网络设计大多都忽视了这一点,直接采用ResNet作为特征提取网络,没有更好地利用多尺度信息. 因此本文提出一种新颖的自适应感受野模块,通过对不同分支的不同感受野特征使用注意力机制使得网络可以自适应地选择合适的感受野特征.

3)为了更加适合实际应用,使用分组卷积替换标准卷积,可减少参数量和计算量,并且网络仅仅使用全局特征,方法简洁高效,在DukeMTMC-reID和Market-1501数据集上均有显著的性能提升.

1 相关工作

近年来,随着卷积神经网络的快速发展,很多有效的卷积方式相继出现. 分组卷积最早由文献[11]提出,可以大大减少模型参数和计算量,获得了广泛的应用. 文献[10]通过空洞卷积在不增加训练参数的情况下增大模型的感受野. InceptionNet在不同的卷积分支上使用不同大小的卷积核,以整合更丰富的多尺度信息.

多尺度特征学习在计算机视觉任务上取得了广泛的成功,在行人重识别领域也发挥着重要作用. 例如,文献[12]提出了一种图像金字塔上提取特征并且共同学习的方法. 与现有的具有多分支模块的行人

重识别模型不同,本文的自适应感受野模块受启发于生物的视觉系统,通过在不同分支设置不同空洞率的空洞卷积来增大感受野,通过对不同分支的多尺度特征使用注意力机制进行更为有效的整合利用.

随着嵌入式AI的发展,轻量级网络的设计引起了广泛的关注. SqueezeNet[13]使用卷积压缩特征通道, IGCNet[14]、ResNeXt[15]和CondenseNet[16]利用分组卷积减少网络参数量和计算量. Xception[17]和MobileNet系列[18-19]进一步采用了深度可分离卷积,取得了更好的效果. 本文提出的自适应感受野模块使用了分组卷积,实现了模型参数量和计算量的降低,模块非常轻量级,适合实际应用场景.

在文献[5]提出的删除特征提取网络中的最后一个空间下采样操作后,行人重识别领域开始广泛采用这一方法,因为在增加特征图的分辨率的同时仅增加非常少量的计算成本和训练参数,但该方法使得网络感受野大大减小,使得网络难以充分利用全局信息.

本文提出的方法仅仅利用全局特征,提出的自适应感受野模块可有效解决行人重识别网络感受野较小的问题. 模块遵循了InceptionNets的思想,为多个分支提供各种不同的卷积方式,区别在于: 1)本文的感受野特征选择模块非常简洁,不需要复杂的定制设计; 2)本文的每个分支都利用了空洞卷积且空洞率各不相同,使得感受野进一步增大,并且采用分组卷积,可实现模块的轻量化; 3)本文使用通道注意力机制,可以让网络自主选择合适的感受野特征,并形成自适应感受野,更好地融合多尺度信息.

2　自适应感受野的行人重识别

行人重识别的全局特征方法比较简洁,无需考虑一些局部信息,直接对整张行人图像进行特征提取. 行人重识别方法最初采用全局特征,但为了追求更好的性能,目前的行人重识别模型往往采用更为复杂的局部特征. 局部特征是指手动或者自动地让网络关注关键的局部区域,然后提取这些区域的行人特征. 常用的提取局部特征的方法有图像切块[20]和关键点定位[21]等. 本节首先介绍一种利用全局特征的基线(baseline)方法,然后阐述所提出的自适应感受野网络的行人重识别方法,包括模型的整体架构以及网络训练和行人重识别过程.

2.1　基线方法

为验证方法的有效性,本文采用强劲的基线方法,其损失函数采用行人ID损失和三元组损失,其特征提取网络采用ResNet-50[22]. 本文使用在ImageNet[23]上预训练的参数初始化ResNet-50,并将全连接层的维度更改为训练数据集中的行人身份数量. 借鉴文献[5, 24, 25],本文对特征提取网络ResNet-50进行调整,没有采用第4阶段中的下采样操作. 通过取消最后的下采样操作,可以得到更大的、细粒度更丰富的特征图,这对于行人重识别是非常重要的.

考虑到行人的比例一般是接近1:2,本文在数据处理部分将图片的宽高比控制在1:2,将输入分辨率调整到384×192. 本文对每张图像都进行概率为0.5的水平翻转和随机擦除、裁剪操作. 需要注意的是,在进行随机裁剪操作之前,先对图片进行0像素填充操作,在图像周围填充10个像素值. 这一系列的数据增广操作可以提高模型的泛化能力并缓解过拟合. 另外,本文也采用标签平滑缓解过拟合. 为了加快模型收敛,本文使用Adam优化器,学习率设置为0.000 35.

2.2　自适应感受野网络方法

本文提出的自适应感受野网络架构如图1所示.

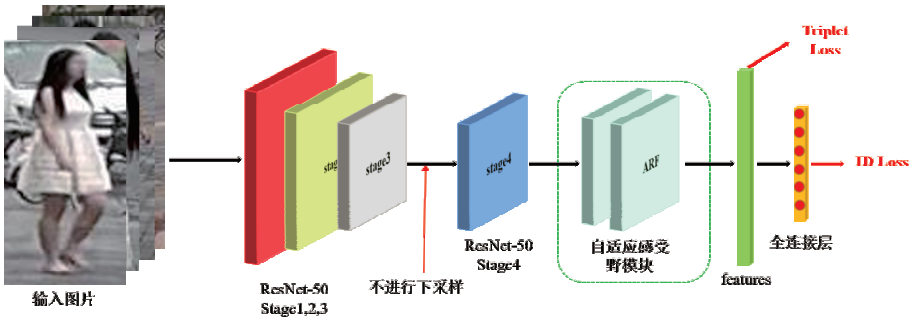


图 1　自适应感受野网络结构

自适应感受野模块可直接应用在特征提取网络末端,操作简便、计算量少. 基线方法(没有自适应感受野模块)的最后阶段没有进行下采样以保持特征图的空间尺寸和特征的细粒度,但容易导致网络的感受野减小. 文献[25]实验发现,当行人重识别采用ResNet-50作为特征提取网络时,输入分辨率如果从256×128增大到384×192,则网络效果反而会下降. 因为删除最后的下采样层,网络的感受野会变小且网络的尺度相对比较单一,无法很好地适应输入分辨率的变化.

本文提出的自适应感受野模块利用空洞卷积和通道注意力机制,弥补了感受野减小带来的不利影响并且自主选择合适的感受野特征,达到了更好的特征提取效果. 自适应感受野模块直接应用在基线网络的末端,其每一个分支都采用不同空洞率的空洞卷积. 空洞卷积可以使网络的感受野显著变大,却没有增加多余的参数量和计算量. 每个分支使用不同的空洞率是为了获得不同大小的感受野,更好地挖掘多尺度信息. 适应感受野模块嵌入到基线网络后,感受野变得更大,便于联合全局信息和聚焦局部区域. 行人重识别领域经常去除特征提取网络最后的下采样,丰富了特征的细粒度但减小了网络的感受野. 本文提出的自适应感受野网络兼顾二者,并且利用注意力机制对多尺度的特征进行了更加有效的整合,使得性能进一步提升.

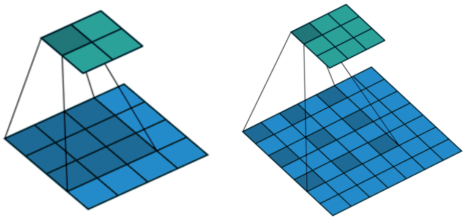
2.3 空洞卷积操作

增大感受野大小是许多视觉任务的关键问题. 在感受野足够大时,网络才可以捕捉更多的图像信息和上下文信息. 然而,有效感受野只是理论感受野的较小部分. 因此,对于大多数视觉任务而言,感受野越大越好. 对于行人重识别任务,网络的感受野大小尤为重要. 理论感受野的计算如下:

$$R_k = R_{k-1} + \left((K_k - 1) \times \prod_{i=1}^{k-1} s_i\right). \tag{1}$$

其中: R_k 表示当前层的感受野大小, R_{k-1} 表示前一层的感受野大小, K_k 表示当前层的卷积核尺寸, s_i 表示第 i 层的步长. 如果取消特征提取网络 ResNet-50 最后的下采样操作,则将 ResNet-50 的第 4 阶段起始步长由 2 设置为 1. 由式(1)可知,网络感受野会大大减小,这对行人重识别非常不利.

图 2 是空洞卷积操作的示意图,绿色特征图代表经过卷积操作输出的特征图,蓝色特征图代表卷积操作输入的特征图.



(a) 空洞率为 1 (b) 空洞率为 2

图 2 空洞卷积操作

图 2(a)是卷积核大小为 3×3、空洞率(dilation)为 1 的空洞卷积操作,属于标准的卷积操作,可以看到输出特征图的每一个点对应输入的一个区域. 图 2(b)是卷积核大小为 3×3、空洞率为 2 的空洞卷积操作,可以看到输出特征图的每一个点对应了输入的一个 5×5 区域. 区域中的阴影部分代表进行操作的点,其余点的权重为 0. 因此,通过空洞率为 2 的空洞卷积操作,使得卷积核由 3×3 的感受野达到了 5×5 的感受野区域,扩大了感受野区域并且不增加额外的参数. 执行空洞卷积操作后的卷积核大小为

$$K'_k = K_k + ((K_k - 1) \times (D - 1)). \tag{2}$$

其中: K'_k 表示经过空洞卷积操作后实际对应的卷积核尺寸, K_k 表示使用的卷积核尺寸, D 表示空洞率.

2.4 自适应感受野模块

更高的空间分辨率能够带来更为丰富的细粒度特征,因此取消特征提取网络最后阶段的下采样操作[5]可以使得特征图空间分辨率大大提高,从而在行人重识别任务上取得显著的效果. 然而这一操作也会有一定的副作用,那就是网络的感受野也会减小.

在卷积神经网络的特征图中,特定位置的特征向量是由前一层固定区域的响应计算得到,此区域就是该特定位置的感受野. 因此,感受野之外的区域不会影响该特定位置的特征向量的大小. 换言之,卷积神经网络仅仅根据感受野区域得到目标信息. 为了弥补感受野减小带来的损失,本文提出一种自适应感受野模块,如图 3 所示. 它利用空洞卷积实现增大感受野的目的,并采用类似 Inception 的多分支结构在不同分支不同空洞率的卷积核,得到不同大小的感受野. 然后,网络通过注意力机制自行选择合适的、不同大小的感受野特征,实现了感受野的自适应,更好地融合利用了多尺度信息.

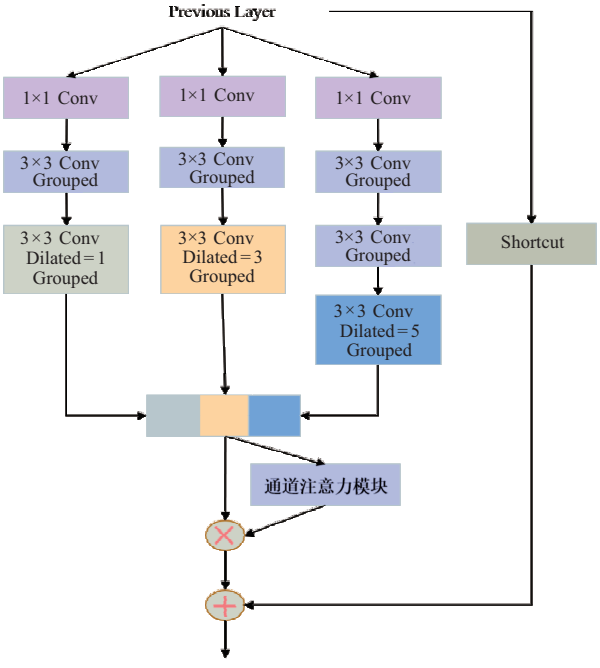


图 3 自适应感受野模块

为了让模块轻量化,本文的自适应感受野模块采用分组卷积进一步减少参数. 本文提出的自适应感受野模块的参数量非常少,是超轻量级的结构,适合实际应用. 当自适应感受野模块应用在ResNet-18网络末端,在分组数group设置为8时,模块参数量只有0.165 M.

神经元的感受野尺寸是不同的,能够在计算时收集多尺度的空间信息. 为了学习到更好的区别性特征,网络还需要提取不同感受野大小的特征并进行有效整合,尤其是当不同行人整体上非常近似时,为了正确进行匹配,有时需要聚焦于图像某一特定区域来排除其他因素的干扰. 然而,目前的行人重识别网络设计大多都忽视了这一点,往往是直接采用ResNet作为特征提取网络,没有更好地利用多尺度信息.

人类的视觉感知系统不会同时处理整个视觉场景,而是有选择性地去聚焦显著的部分. 注意力在人类的视觉感知中起到了非常重要的作用,注意力模型在很多深度学习任务中有广泛的应用. 基于深度学习的注意力机制与人类的视觉注意力类似,核心目的是抽取重要的信息. 本文利用通道注意力SENet从不同感受野特征中进行选择,使得网络可以实现感受野的自适应. 通道注意力机制结构如图4所示,不同感受野的多尺度特征通过在不同通道乘以通过右侧分支学习到的权重系数来动态地聚合多尺度信息.

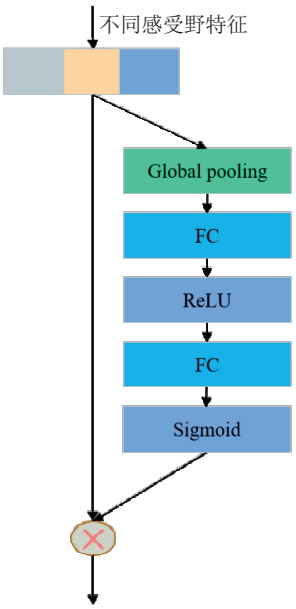


图 4　通道注意力模块

具体而言,右侧分支先是通过全局平均池化整合各个通道的信息,再通过两个全连接层去建模通道间的相关性,然后通过Sigmoid函数得到归一化后的权重,最后将归一化后的权重加权到每个通道的特征上. 模块利用注意力机制对不同大小感受野的多尺度特征进行了有效的自适应选择,使得感受野更加符合实际,与生物的视觉系统更为相近. 例如,当两个人整体外观上非常近似的时候,自适应感受野网络就会更加聚焦于一些局部区域的差别(如鞋子、发型),关注这些局部区域与关注整体全局外观都是非常重要的,而本文提出的自适应感受野网络模块可以更好地整合不同感受野特征,充分利用了多尺度信息,可以取得非常好的效果.

3　实验结果与分析

为了验证本文算法的效果,本文在两个大型公开数据集Market-1501和DukeMTMC-reID上开展实验测试.

3.1　实验数据集和评价指标

DukeMTMC-reID数据集由8个不同的摄像头采集的图片组成,总共含有1 404个不同行人共计36 411张图片. 图片是由人工标注完成的,数据集提供了训练集和测试集. 训练集中总共有16 522张图片,包括了702个不同的行人,每个行人平均有23.5张训练图片.

Market-1501数据集是6个不同的摄像头采集的图片,其中一个摄像头分辨率较低. Market-1501数据集包含了1 501个不同身份的行人,一共含有32 668张行人图片. 图片是由检测器自动检测出来的,因此包含了一些检测误差(更加接近实际). Market-1501数据集的训练集包含了751个不同身份行人的12 936张裁剪好的图片. 训练集的每个身份的人平均有17.2张图片. 测试集则包含了750个不同身份的19 732张裁剪好的图片.

3.2　实验设置及评价标准

本文实验是基于Pytorch深度学习框架实现的,所采用的硬件为配备高性能显卡Nvidia GTX Titan Xp的服务器. 对于模型的性能评估,采用了行人重识别领域广泛采用的评价标准,一选准确率(Rank-1)和平均准确度(mean average precision, mAP). Rank-1表达了应当被查询出来的图片在候选序列第1个位置的概率,mAP则反映了这个行人重识别算法模型的总体性能.

3.3　模型评估结果

在DukeMTMC-reID数据集上,使用基线方法和本文提出的自适应感受野网络的实验结果对比如表1所示. 实验结果表明,本文提出的自适应感受野网络在基线方法基础上仅仅增加了两个轻量级的自适应感受野模块,效果却得到了显著的提升. 从表1实验数据可以看到,Rank-1提高了3.2 %,mAP提高了

0.4 %. 本文采取的基线是非常强劲的,Rank-1和mAP已经非常出色,但本文方法仍然取得了显著的效果,表明本文提出的自适应感受野模块可以让网络学习到多尺度且更有分辨性的特征,达到更好的行人重识别效果. 但是添加了本文的模块后,网络的参数量和计算量都会有小幅度的提升,牺牲了较少的时间,但是达到了非常高的精度.

表 1 DukeMTMC-reID数据集上基线方法与本文方法的比较 %

方法	Rank-1	mAP
基线(ResNet-50)	86.0	75.6
本文(ResNet-50)	**89.2**	**89.2**

在Market-1501数据集上,使用基线方法和本文提出的自适应感受野网络的实验结果如表2所示,本文提出的自适应感受野网络在基线方法基础上得到了显著的提升,其中Rank-1提高了1.3 %,mAP提高了1.8 %.

表 2 Market-1 501数据集上基线方法与本文方法的比较 %

方法	Rank-1	mAP
基线(ResNet-50)	86.0	75.6
本文(ResNet-50)	**89.2**	**76.0**

3.4 使用不同的特征提取网络的测试结果

为了进一步凸显自适应感受野模块的有效性和通用性,本文采用不同的经典模型作为特征提取网络进行对比实验,在DukeMTMC-reID数据集和Market-1 501数据集上得到的结果如表3和表4所示.

表 3 DukeMTMC-reID数据集上基线方法与本文方法采用不同特征提取网络的比较 %

方法	Rank-1	mAP
基线(Mobilenetv2)	83.8	69.8
本文(Mobilenetv2)	85.3	71.9
基线(ResNet-18)	85.2	72.0
本文(ResNet-18)	87.3	74.3
基线(ResNet-34)	86.4	75.1
本文(ResNet-34)	**88.6**	**77.8**
基线(SENet-50)	85.5	75.4
本文(SENet-50)	87.4	77.2

表3的实验结果表明,本文的自适应感受野网络是非常实用和通用的. 本文提出的方法在使用ResNet-18作为特征提取网络时的性能已经超过基线使用ResNet-34作为特征提取网络的结果,表明所提出的自适应感受野网络在更少的参数量情况下实现了更好的性能,可以更好地满足行人重识别实际应用的需求. 当采用性能较弱的轻量级网络Mobilenetv2和ResNet-18时,在两个数据集上mAP的提高尤为显著,表明本文算法对于网络特征提取能力的加强是非常有效的.

表 4 Market-1 501数据集上基线方法与本文方法采用不同特征提取网络的比较 %

方法	Rank-1	mAP
基线(Mobilenetv2)	92.2	79.0
本文(Mobilenetv2)	93.5	81.4
基线(ResNet-18)	92.0	79.8
本文(ResNet-18)	94.3	85.2
基线(ResNet-34)	93.2	84.2
本文(ResNet-34)	94.5	86.9
基线(SENet-50)	94.7	85.6
本文(SENet-50)	**87.4**	**87.1**

3.5 本文方法与基线方法的参数量比较

本文提出的自适应感受野模块是非常轻量级的,可以简便地应用到现有的特征提取网络. 基线方法使用不同特征提取网络的参数量与自适应感受野模块参数量的对比如表5所示.

表 5 基线方法与本文方法的参数量比较

方法	#P(M)
基线(ResNet-50)	25.05
本文(ResNet-50)	25.45

从表5数据可知,本文模块在ResNet-50中所占用的参数量只有0.4 M,模块非常轻量级. 由于整个自适应感受野模块的内部网络参数量主要随着通道数目变化而变化,与特征提取网络息息相关,如果特征提取网络是ResNet-50,则本文模块的输入通道数就是2 048,此时单个模块的参数量为0.2 M.

3.6 本文方法与其他先进的行人重识别算法比较

本文算法与4种利用局部特征的方法(AlignedReID[21](aligned re-identification)、SCPNet[26](spatial-channel parallelism network)、PCB[5](part-based conveolutional baseline)、BFE[27](batch feature erasing)),2种利用全局特征的方法(SVDNet[28](singular vector decomposition network)、AWTL[29](adaptive weighted triplet loss))以及先进基线方法在DukeMTMC-reID数据集和Market-1501数据集上进行了实验对比. 本文方法与先进方法的比较如表6所示.

表 6　与先进方法的比较

类型	方法	DukeMTMC-reID		Market-1501	
		Rank-1	mAP	Rank-1	mAP
局部特征	AlignedReID	81.2	67.4	90.6	77.7
	SCPNet	80.3	62.6	91.2	75.2
	PCB	83.3	69.2	93.8	81.6
	BFE	88.7	75.8	94.5	85.0
全局特征	SVDNet	76.7	56.8	82.3	62.1
	AWTL	79.8	63.4	89.5	75.7
	基线方法	87.2	74.2	93.8	85.3
	本文方法	**89.2**	**76.0**	**95.2**	**87.2**

从表6数据可以看到,当前效果较好的算法大多采用了局部特征,但模型变得复杂并且实用性降低. 本文从实用性考虑,采用了强劲的全局特征模型作为基线并进行改进. 从表6可以看出,本文提出的自适应感受野网络超过了其他对比算法,验证了本文方法的有效性. BFE算法[27]与本文算法有相近的性能表现,但它结合了两个分支的特征,包括全局特征和局部特征,这会导致网络变得更加复杂,而本文仅利用全局特征就取得了非常先进的性能. 多尺度信息在许多其他视觉任务上都取得了非常好的效果,本文受启发于生物的视觉系统,设计了自适应感受野模块来更有效地在行人重识别任务上结合不同尺度特征. 本文算法在实际场景的图片上进行特定行人检索的结果如图5所示.

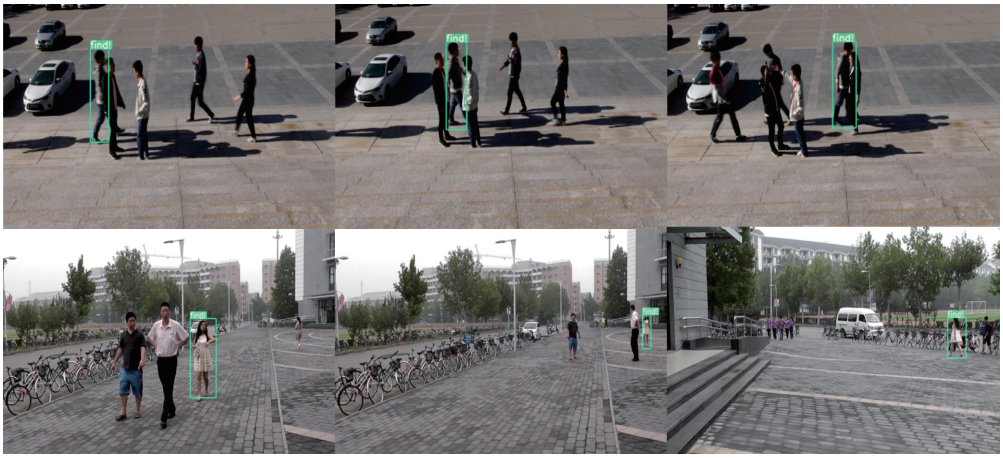


图 5　实际场景图片进行特定行人检索的结果

由图5可以看到,在一些遮挡情况以及行人尺度有较大变化的情况下均能识别出特定的行人,验证了本文算法在实际场景的有效性.

4　结　论

现有行人重识别方法为了保留更多的细粒度特征,通常删除特征提取网络中的最后一个空间下采样操作,以增加最后输出特征图的分辨率. 然而,这种操作会大幅减小神经网络的感受野,而更大的感受野可以为行人重识别提供更多的上下文信息. 删除特征提取网络中最后一个空间下采样操作会造成感受野减小,本文针对此问题提出了一种轻量级的自适应感受野模块来弥补感受野的减小,并自动选择不同大小的感受野得到多尺度且具有辨别性的特征. 相比局部特征,全局特征更适合实际应用,本文将自适应感受野模块应用在基于全局特征的强劲行人重识别基线上,在DukeMTMC-reID和Market-1501数据集上进行了实验. 实验结果表明,本文的自适应感受野模块可以显著改善行人重识别的性能,简单有效,非常适合实际应用.

参考文献(References)

[1] Zhu F Q, Kong X W, Fu H Y, et al. Two-stream complementary symmetrical CNN architecture for person re-identification[J]. Journal of Image and Graphics, 2018, 23(7): 1052-1060.

[2] Ben X Y, Xu S, Wang K J. Review on pedestrian gait feature expression and recognition[J]. Pattern Recognition and Artificial Intelligence, 2012, 25(1): 71-81.

[3] Yi D, Lei Z, Liao S C, et al. Deep metric learning for person re-identification[C]. Proceedings of the 22nd International Conference on Pattern Recognition. Stockholmo: IEEE, 2014: 34-39.

[4] Li W, Zhao R, Xiao T, et al. DeepReID: Deep filter

pairing neural network for person re-identification[C]. Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus: IEEE, 2014: 152-159.

[5] Sun Y F, Zheng L, Yang Y, et al. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline)[C]. Proceedings of the European Conference on Computer Vision (ECCV). Munich: Springer, 2018: 480-496.

[6] Hubel D H, Wiesel T N. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex[J]. The Journal of Physiology, 1962, 160(1): 106-154.

[7] LeCun Y, Boser B, Denker J S, et al. Backpropagation applied to handwritten zip code recognition[J]. Neural Computation, 1989, 1(4): 541-551.

[8] Szegedy C, Liu W, Jia Y Q, et al. Going deeper with convolutions[C]. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston: IEEE, IEEE, 2015: 1-9.

[9] Hu J, Shen L, Sun G. Squeeze-and-excitation networks[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake: IEEE, 2018: 7132-7141.

[10] Chen L C, Papandreou G, Kokkinos I, et al. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(4): 834-848.

[11] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[C]. Proceedings of the 25th International Conference on Neural Information Processing Systems. Lake Tahoe: Curran Associates Inc., 2012: 1097-1105.

[12] Chen Y B, Zhu X T, Gong S G. Person re-identification by deep learning multi-scale representations[C]. Proceedings of 2017 IEEE International Conference on Computer Vision Workshops. Venice: IEEE, 2017: 2590-2600.

[13] Iandola F N, Han S, Moskewicz M W, et al. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and 0.5 MB model size[EB/OL]. (2016-11-04)[2019-12-02]. https://arxiv.org/abs/1602.07360.

[14] Zhang T, Qi G J, Xiao B, et al. Interleaved group convolutions[C]. Proceedings of the IEEE International Conference on Computer Vision. Venice: IEEE, 2017: 4373-4382.

[15] Xie S, Girshick R, Dollár P, et al. Aggregated residual transformations for deep neural networks[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017: 1492-1500.

[16] Huang G, Liu S C, Maaten L V D, et al. Condensenet: An efficient densenet using learned group convolutions[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Sale Lake: IEEE, 2018: 2752-2761.

[17] Chollet F. Xception: Deep learning with depthwise separable convolutions[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE: 2017: 1251-1258.

[18] Howard A G, Zhu M, Chen B, et al. Mobilenets: Efficient convolutional neural networks for mobile vision applications[EB/OL]. (2017-04-17)[2019-12-02]. https://arxiv.org/abs/1704.04861.

[19] Sandler M, Howard A, Zhu M, et al. Mobilenetv2: Inverted residuals and linear bottlenecks[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake: IEEE, 2018: 4510-4520.

[20] Varior R R, Haloi M, Wang G. Gated siamese convolutional neural network architecture for human re-identification[J]. European Conference on Computer Vision, DOI:10.10071978-3-319-46484-8.48.

[21] Zhang X, Luo H, Fan X, et al. AlignedReID: Surpassing human-Level performance in person re-Identification[EB/OL]. (2018-01-31)[2019-12-02]. https://arxiv. org/abs/1711.08184.pdf.

[22] He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition[C]. Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016: 770-778.

[23] Deng J, Dong W, Socher R, et al. Imagenet: A large-scale hierarchical image database[C]. Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Miami: IEEE, 2009: 248-255.

[24] Wang G S, Yuan Y F, Chen X, et al. Learning discriminative features with multiple granularities for person re-identification[C]. Proceedings of the Conference on Multimedia. New York: ACM, 2018: 274-282.

[25] Luo H, Gu Y Z, Liao X Y, et al. Bag of tricks and a strong baseline for deep person re-identification[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. Long Beach: IEEE, 2019: 1487-1495.

[26] Fan X, Luo H, Zhang X, et al. Scpnet: Spatial-channel parallelism network for joint holistic and partial person re-identification[C]. Asian Conference on Computer Vision. Perth: Springer, 2018: 19-34.

[27] Dai Z, Chen M, Zhu S, et al. Batch feature erasing for person re-identification and beyond[EB/OL]. (2018-11-17)[2019-07-20]. https://arxiv.org/abs/1811.07130.pdf.

[28] Sun Y F, Zheng L, Deng W J, et al. SVDNet for pedestrian retrieval[C]. IEEE International Conference on Computer Vision. Venice: IEEE, 2017: 3820-3828.

[29] Ristani E, Tomasi C. Features for multi-target multi-camera tracking and re-identification[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake: IEEE, 2018: 6036-6046.

作者简介

王松 (1995−), 男, 硕士生, 从事计算机视觉领域的研究, E-mail: wangsong@stumail.neu.edu.cn;

纪鹏 (1980−), 男, 讲师, 从事人工智能与模式识别领域等研究, E-mail: jipeng@mail.neu.edu.cn;

张云洲 (1974−), 男, 教授, 博士, 从事智能机器人、计算机视觉等研究, E-mail: zhangyunzhou@mail.neu.edu.cn;

朱尚栋 (1992−), 男, 博士生, 从事计算机视觉和行人重识别、机器学习的研究, E-mail: zhushangdong@gmail.com;

暴吉宁 (1990−), 女, 博士生, 从事计算机视觉和动态目标跟踪的研究, E-mail: yiyinbaobao@126.com.

(责任编辑：闫 妍)