控制与决策
Control and Decision

基于多尺度残差注意网络的轻量级行人属性识别算法

张再腾, 张荣芬, 刘宇红

引用本文:
张再腾,张荣芬,刘宇红. 基于多尺度残差注意网络的轻量级行人属性识别算法[J]. 控制与决策, 2022, 37(10): 2487−2496.

在线阅读 View online: https://doi.org/10.13195/j.kzyjc.2021.0411

基于多尺度残差注意网络的轻量级行人属性识别算法

张再腾, 张荣芬, 刘宇红†

(贵州大学 大数据与信息工程学院，贵阳 550025)

摘 要: 近年来,随着深度学习的蓬勃发展,行人属性识别得到了广泛的研究. 但是,由于属性复杂且多样化、图像质量差、视角遮挡等困扰,难以捕获图像中的细粒度属性特征,具有很大的挑战性. 对此,基于深度学习,提出多尺度残差注意网络(MRAN)用于行人属性识别,以Resnet 50为主体架构,使用轻量级的金字塔卷积提供不同内核大小的并行卷积以完成多尺度信息的提取,嵌入注意力模块以关注属性存在的关键区域并挖掘属性内部联系;其次,使用特征金字塔融合策略,更充分地提取和融合多尺度特征. 网络结合了多尺度学习、注意力机制和残差学习的思想,使网络提取出更丰富、更细腻的特征. 最后,在PETA和PA100K两个数据集上进行实验研究,结果表明,所提出方法优于现有的研究方法. 通过消融研究验证整个网络体系结构的3个组成部分的有效性和先进性,且所提出网络具有高准确性和低复杂度的双向优化.

关键词: 行人属性识别；多尺度；金字塔卷积；注意力机制；特征金字塔；轻量级

中图分类号: TP391.41 **文献标志码:** A

DOI: 10.13195/j.kzyjc.2021.0411 **开放科学(资源服务)标识码(OSID):**

引用格式: 张再腾,张荣芬,刘宇红. 基于多尺度残差注意网络的轻量级行人属性识别算法[J]. 控制与决策, 2022, 37(10): 2487-2496.

Lightweight pedestrian attribute recognition algorithm based on multi-scale residual attention network

ZHANG Zai-teng, ZHANG Rong-fen, LIU Yu-hong†

(College of Big Data and Information Engineering，Guizhou University，Guiyang 550025，China)

Abstract: Recently, pedestrian attribute recognition has been extensively studied that has benefited from the vigorous development of deep learning. However, it is difficult to capture the fine-grained attributes in the image due to complex and diversified attributes, poor image quality, and viewing angle occlusion, which is very challenging. Based on deep learning, we propose a multi-scale residual attention network (MRAN) for pedestrian attribute recognition with Resnet 50 as the main architecture, using lightweight pyramid convolution to provide parallel convolution with different kernel sizes to complete multi-scale information extraction. The attention module is embedded to focus on the key areas where the attributes exist and explore the internal relations of the attributes. Secondly, the feature pyramid aggregation strategy is used to more fully extract and fuse multi-scale features. The network combines the ideas of multi-scale learning, attention mechanism and residual learning to enable the network to extract richer and more delicate features. Finally, an experimental study is carried out on the two datasets of PETA and PA100K, and the results show that the proposed method is superior to the existing research methods. Through ablation research, the effectiveness and advancement of the three components of the entire network architecture are verified, and the proposed network has bidirectional optimization with high accuracy and low complexity.

Keywords: pedestrian attribute recognition；multi-scale；pyramid convolution；attention mechanism；feature pyramid；lightweight

0 引 言

行人属性识别[1]旨在给定人物图像或视频序列时挖掘目标人物的属性特征,行人属性包括行人的性别、年龄、着装、配饰等信息,其具有通过图像信息提取语义特征来描述行人在一个场景中的特点. 在计算机视觉的快速发展下,通过深度学习提取特征的方法逐渐取代了传统手工获取特征的方法,基于深度学习的行人属性识别得到了蓬勃的发展,被广泛应用

收稿日期: 2021-03-10；**录用日期:** 2021-07-05.

基金项目: 贵州省科学技术基金项目(黔科合基础-ZK[2021]重点001).

†通讯作者. E-mail: liuyuhongyx@sina.com.

于智能监控、视频图像检索以及人员重识别等领域,在智能安防中展现出了巨大的潜力.

自里程式的LeNet[2]、AlexNet[3]等卷积神经网络提出后,计算机视觉得到了极速发展且展现了出色的性能. 与此同时,行人属性识别也在卷积神经网络的依托下脱离了SVM[4]、SIFT[5]等传统提取特征的方法. 许多学者对基于深度学习的行人属性识别进行了深入研究,已经提出了许多算法来提升行人属性识别性能. 为了获取属性的多尺度细节特征,基于注意力的模式,人们提出了HPnet[6]、VAC[7]、CAS[8]等模型,该类方法虽然关注了区域中属性特征,但忽略了样本属性内部联系的影响. 对于属性的相关性研究大多使用CNN+RNN (convolutional neural network +recurrent neural network)模型,如: JRL[9]、GRL[10]、MTA-Net[11]. 最近还提出了使用GCN (graph convolutional network)来关注属性关系,通过扩展神经网络,使用图结构来处理数据,以捕获属性关系的上下文关系,有VSGR[12]、CGCN[13]等方法. 这种复杂的注意力或利用CNN-RNN、GCN关注属性的方法虽然有效提升了性能,但极大地增加了参数的数量,同时降低了识别速度,而且图像中人物属性间的复杂关系和属性目标大小分布不平衡问题并没有得到很好地解决.

针对目前行人属性识别数据集分辨率不均、角度多变化等因素导致的属性大小分布不均衡以及属性间的相关性问题,本文提出一种多尺度残差注意网络的轻量级行人属性识别模型,通过多尺度残差的方式来捕获小目标属性细节,融入注意力模型以挖掘属性之间的深层消息. 该模型端到端地学习全局特征并细腻化局部特征,获取更丰富的特征表示,提升整体属性识别能力. 本文主要做了以下工作:

1) 基于Resnet 50[14]框架采取金字塔卷积及特征金字塔策略,增加图像感受野,以多种尺度解析,捕捉更细腻的信息,提高小目标属性识别能力,提升整体模型性能.

2) 由于行人位于图像的不同位置,其所属属性也分布于不同区域,为了有效关注属性所在的区域,本文采取新颖的注意力方式,该方式能有效融合通道与空间信息,增加属性之间的相关性.

3) 在实际应用中,模型的大小和运算速度也是考虑的重点,一般识别精度和模型大小不能达到较好的平衡,而本文采取的金字塔卷积是基于分组卷积和金字塔形式的卷积方式,其组合了不同的内核类型,能同时提取深层和浅层特征,且具有较少参数量和计算量,从内部减少了模型的存储和计算成本.

1 相关工作

由于深度学习在特征学习中展现出的有效性,许多研究人员利用深度学习来解决行人属性识别问题. 总体而言,基于深度学习的行人属性识别算法可分为以下4类:

1) 基于全局的模型. 文献[15]提出ACN (attribute convolutional net),通过联合训练整体CNN模型来学习不同的属性; DeepMAR[16]将多个属性进行联合,实现多属性分类.

2) 基于局部的模型. PGDM[17]考虑行人动作和姿势关键点来提取局部区域,将局部区域与整体特征相融合进行属性识别; LGNet[18]根据预先提取的区域与属性位置之间的相似性,将属性特定的权重分配给局部特征; 文献[19]提出了一种属性本地化模块,实现自动划分区域并在局部区域内提取属性特征.

3) 基于注意力的模型. MsVAA[20]在不同的尺度上提取和聚集视觉注意任务; CAS[8]提出共同的方式共享模块,以软功能共享两个网络,实现人的属性识别; HPnet[6]和VAC[7]也属于此类.

4) 基于顺序预测的模型. JRL[9]探索属性与视觉环境之间的相互依赖性和相关性; 类似的方法还有GRL[10]、MTA-Net[11]、RCRA[21]等.

最近,还有许多其他方法,比如: 使用图卷积方式; PD-Net[22]提出利用提炼分支和属性融合分支来挖掘属性相关性; MCFL[23]改善损失函数以解决属性数量分布不均衡问题; 文献[24]引入Gabor小波结合CNN解决行人属性识别问题.

虽然基于深度学习的方法在不同的模型上展现了有效的性能,但也存在相对的局限性. 第1类方法虽然简单直观,但是缺乏对特征细腻度识别的考虑,导致识别模型的性能受到限制; 第2类使用局部信息能显著提高整体识别性能,但是,不正确的局部检测结果将会导致最终分类的错误; 第3类现有的注意力机制没有关注到属性之间的相对关联性,使得效果有待提升; 第4类基于顺序估计的模型虽然有效,但是,由于连续的属性估计会导致时间效率的低下,不利于实际应用.

基于以上问题,本文设计一种多尺度残差注意的框架处理行人属性识别问题,以Resnet 50为主体框架,内置金字塔卷积与注意力模块构成PCA-Resnet 50 (PyConv-attention-Resnet 50),以提升特征细腻度识别和属性的区域关注度. 网络后端使用特征金字塔进行融合,将浅层小目标的位置信息与深层语义信息进行整合,进而提升整体性能. 相比于原始

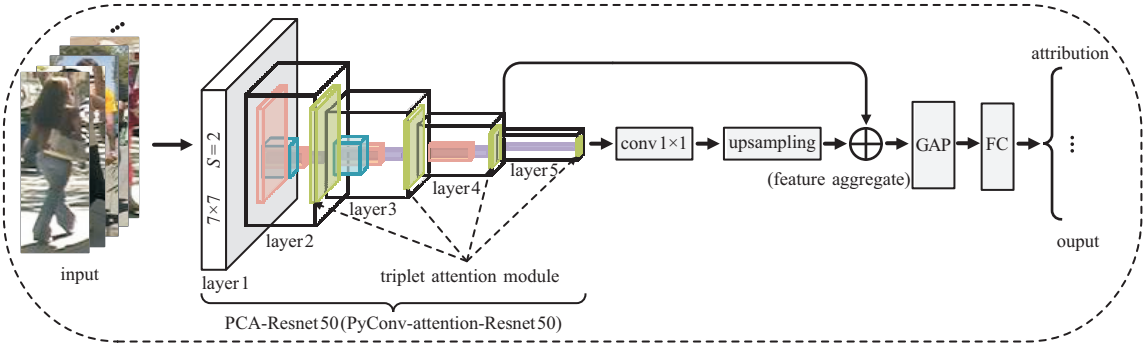


图 1　multi-scale residual attention network (MRAN) 总体结构

Resnet 50, 准确度更高, 复杂度更低, 是一种端到端的轻量架构. 网络框架结构如图 1 所示.

2　提出算法

本文提出一种基于多尺度残差注意的轻量级行人属性识别模型, 基于 Resnet 50 网络架构插入注意力模块以增强对图像属性所在区域的识别. 该网络不仅使用了金字塔卷积, 还应用了特征金字塔策略, 此策略可以整合浅层与深层之间的特征, 以增强对小目标属性的识别, 进而提高整体识别效果.

2.1　金字塔卷积

卷积神经网络的核心是卷积核, 大多数神经网络使用相对较小的内核, 通常为 3×3. 由于增加内核大小会使得参数数量和计算复杂性的成本增加, 且标准卷积只具有单一空间大小的单一类型的核, 并不具备在多个尺度上处理输入的能力, 由此逐渐衍生出空洞卷积[25]、深度可分离卷积[26]、分组卷积[3] 等卷积方式. 金字塔卷积 (pyramidal convolution, PyConv)[27] 与标准卷积相比它包含了不同大小和深度的不同级别的核. 除了扩大感受野之外, PyConv 还可以使用并行增加的内核大小来处理输入, 可以更好地捕获多尺度的细节信息.

图 2(a) 所示的标准卷积包含单一类型的核: 具有单一空间大小 K_1^2, 标准卷积所需的参数为 $K_1^2 \cdot \mathrm{FM}_i \cdot \mathrm{FM}_o$, FLOPs (浮点运算) 数为 $K_1^2 \cdot \mathrm{FM}_i \cdot \mathrm{FM}_o \cdot (W \cdot H)$. 其中: FM_i 表示输入特征图, FM_o 表示输出特征图, H 为高, W 为宽. 为了能够在 PyConv 的每个级别使用不同深度的核, 利用分组卷积的方式将输入特征映射分成不同的组, 并为每个输入特征映射组独立地应用内核.

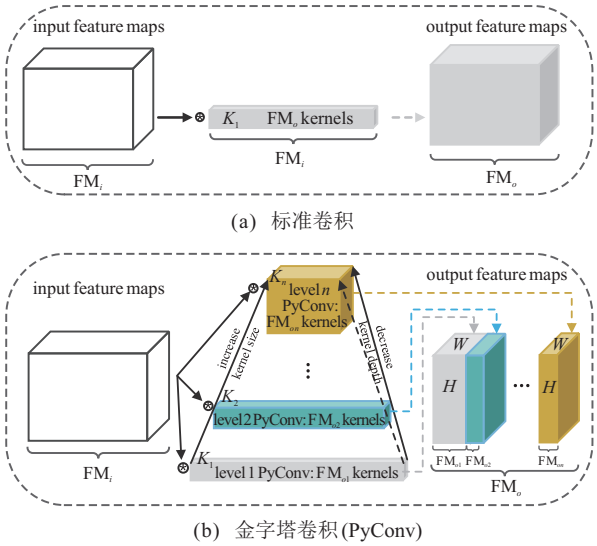


图 2　金字塔卷积

如图 3(a) 所示, 当 G (groups) $= 1$ 时为标准卷积, 此时每个输出要素图都连接到所有输入要素图. 图 3(b) 显示了 $G = 2$ 时将输入特征映射分成两组的情况, 其中核被独立地应用于每组, 因此, 核的深度被减少了 2 倍. 图 3(c) 显示当 $G = 4$ 时, 核的深度减少了 4 倍. 当组数量增加时, 连通性和核的深度均降低, 使得卷积的参数数量和计算成本减少.

如图 2(b) 所示, 对于输入特征图 FM_i, 金字塔卷积的 $\{1, 2, \ldots, n\}$ 每个级别对应于不同的空间大小内核 $\{K_1^2, K_2^2, \ldots, K_n^2\}$, 通过图解中分组的方式得到不同深度的核为

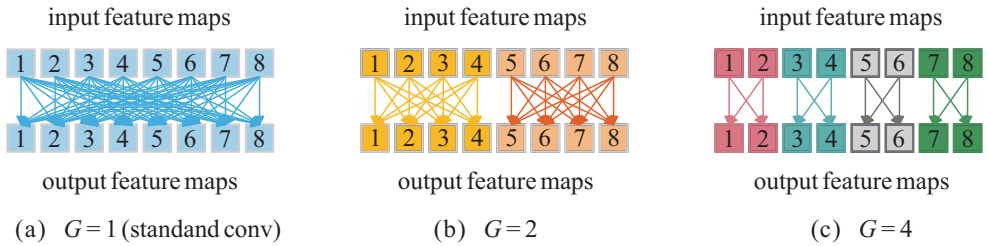


图 3　分组卷积

$$\left\{\mathrm{FM}_i, \frac{\mathrm{FM}_i}{\left(\frac{K_2^2}{K_1^2}\right)}, \frac{\mathrm{FM}_i}{\left(\frac{K_3^2}{K_1^2}\right)}, \ldots, \frac{\mathrm{FM}_i}{\left(\frac{K_n^2}{K_1^2}\right)}\right\}.$$

其所需的参数为

$$K_n^2 \cdot \frac{\mathrm{FM}_i}{\left(\frac{K_n^2}{K_1^2}\right)} \cdot \mathrm{FM}_{on} + \ldots + K_3^2 \cdot \frac{\mathrm{FM}_i}{\left(\frac{K_3^2}{K_1^2}\right)} \cdot \mathrm{FM}_{o3} + K_2^2 \cdot \frac{\mathrm{FM}_i}{\left(\frac{K_2^2}{K_1^2}\right)} \cdot \mathrm{FM}_{o2} + K_1^2 \cdot \mathrm{FM}_i \cdot \mathrm{FM}_{o1}; \tag{1}$$

FLOPs数为

$$K_n^2 \cdot \frac{\mathrm{FM}_i}{\left(\frac{K_n^2}{K_1^2}\right)} \cdot \mathrm{FM}_{on} \cdot (W \cdot H) + \ldots + K_3^2 \cdot \frac{\mathrm{FM}_i}{\left(\frac{K_3^2}{K_1^2}\right)} \cdot \mathrm{FM}_{o3} \cdot (W \cdot H) + K_2^2 \cdot \frac{\mathrm{FM}_i}{\left(\frac{K_2^2}{K_1^2}\right)} \cdot \mathrm{FM}_{o2} \cdot (W \cdot H) + K_1^2 \cdot \mathrm{FM}_i \cdot \mathrm{FM}_{o1} \cdot (W \cdot H). \tag{2}$$

其中: 输出特征图为 $\{\mathrm{FM}_{o1}, \mathrm{FM}_{o2}, \ldots, \mathrm{FM}_{on}\}$, 且 $\mathrm{FM}_{o1} + \mathrm{FM}_{o2} + \ldots + \mathrm{FM}_{on} = \mathrm{FM}_o$, 即每一级特征图按通道连接得到输出特征图.

金字塔卷积的内核类型是双向金字塔: 一边的内核大小在增加, 另一边的内核深度从级别1到级别n减少; 反之亦然. PyConv不同类型的内核带来了互补的信息, 有较小感受野的内核可以关注细节, 捕捉较小对象的信息; 而增加内核大小, 可以提供较大对象更可靠的细节, 网络在学习过程中不断探索. 该方式可以从连通性较低的大感受野探索到连通性较高的较小感受野, 实现了以不同的内核规模处理输入, 在减少计算成本和模型复杂性的情况下提高了识别精度, 同时使并行性也得到了提高.

2.2 注意力机制

行人属性具有多样性, 行人的尺度、姿态和属性的类内差异很大, 并且属性在图像上的具体位置是随机变化的. 所以, 行人多属性识别的关键是如何使模型关注属性存在的位置信息. 近年来, 不断更迭的注意力模型如SEnet[28]、CBAM[29], 使得网络有目标地学习, 进一步关注目标对象区域, 提升了识别准确率. 行人属性识别也得到了注意力的启发与助力, 虽然CBAM使用空间注意力和通道注意力来显示性能的提高, 但是, 其缺点是通道注意与空间注意是分离的, 并且彼此独立地计算, 没有考虑两者之间的联系. 而行人属性在通道和空间上是有一定程度联系的, 例如: 眼镜位于头部, 背包位于背部等. 因此, 引入新的注意模型triplet attention[30], 使属性分类模型关注行人图像中存在属性的关键区域, 以提高多属性识别的准确性.

triplet attention引入了跨维相互作用, 即通过3个分支分别捕获输入张量的(C, H)、(C, W)和(H, W)之间的依赖关系. 其中: 两个分支负责捕捉通道维度C和空间维度H或W之间的跨维度相互作用, 跨维度相互作用是通过旋转操作来建立通道维度与空间维度的连接; 第3个分支用于捕获空间相关性(H和W), 然后将3个分支的所有输出进行汇总, 计算过程如图4所示.

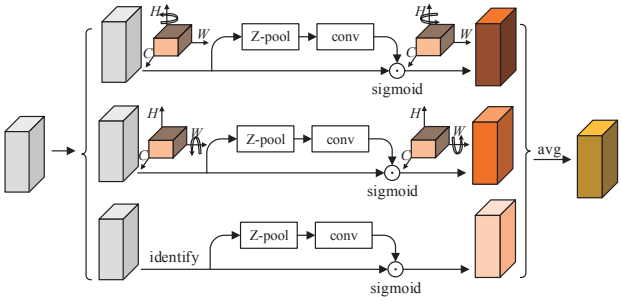


图4 triplet attention

输入一个张量$\chi \in \boldsymbol{R}^{C\times H\times W}$, 首先将输入$\chi$给每个分支. 在第1个分支中, 构建高度维度$(H)$与通道维度$(C)$之间的交互. 将输入$\chi$沿$H$轴逆时针旋转90°后得到$\hat{\chi}_1$的大小$(W \times H \times C)$; 然后, Z-pool通过拼接平均池化的特征以及最大池化的特征, 将张量的第零维度的通道数缩减为2, 即$\hat{\chi}_1$通过Z-pool后得到大小为$(2\times H\times C)$的$\hat{\chi}_1^*$; 随后通过内核大小为$K\times K$的标准卷积层和批量归一化层, 得到提供维度的中间输出$(1 \times H \times C)$. 张量穿过sigmoid激活层(σ)生成最终的注意力权重ω_1. 生成的注意力权重随后被应用于$\hat{\chi}_1$, 然后再沿着H轴顺时针旋转90°以保持χ的原始输入形状. 第2个分支构建宽度维度(W)与通道维度(C)之间的交互, 将输入χ沿W轴逆时针旋转90°, 进行与第1个分支相同的工作, 得到相应的$\hat{\chi}_2$、$\hat{\chi}_2^*$、ω_2. 最后一个分支中输入张量χ, 通过Z-pool后得到$\hat{\chi}_3$的大小为$(2 \times H \times W)$, 经过内核为K的标准卷积和批量归一化层操作后得到$(1 \times H \times W)$, 最后, 经sigmoid激活后得到注意力权重ω_3, 并应用于输入χ. 最终聚集3个分支中的每一个生成张量, 再平均化得到$(C \times H \times W)$大小的triplet attention. 计算公式如下:

$$Z\text{-}\mathrm{pool}(\chi) = [\mathrm{maxpool}_{0d}(\chi), \mathrm{avgpool}_{0d}(\chi)], \tag{3}$$

其中$0d$是进行最大池化和平均池化操作的第0维.

$$y=\frac{1}{3}(\overline{\hat{\chi}_1\sigma\psi_1(\hat{\chi}_1^*)}+\overline{\hat{\chi}_2\sigma\psi_2(\hat{\chi}_2^*)}+\chi\sigma\psi_3(\hat{\chi}_3)). \tag{4}$$

其中: σ代表sigmoid激活函数, ψ_1、ψ_2和ψ_3代表3个分支中由核大小K定义的标准二维卷积层.

简化后得到

$$y=\frac{1}{3}(\overline{\hat{\chi}_1\omega_1}+\overline{\hat{\chi}_2\omega_2}+\chi\omega_3)=\frac{1}{3}(\overline{y_1}+\overline{y_2}+y_3). \tag{5}$$

其中: ω_1、ω_2和ω_3是计算的3个交叉维度的注意力权重, $\overline{y_1}$和$\overline{y_2}$表示90°顺时针旋转以保持$(C\times H\times W)$的原始输入形状.

2.3　多尺度特征金字塔融合

一些小目标属性, 如眼镜、鞋子等, 只具有很小的图像分辨率, 因此, 多属性识别任务的另一个问题是如何提高小目标属性的识别精度. 原始的Resnet 50网络很深, 会丢失小目标的位置信息, 使得多属性识别中难以获取小目标的特征信息. 针对这一问题, 不仅在Resnet 50网络剩余模块中使用金字塔卷积, 而且在整体结构后端引入特征金字塔策略, 将表层小目标的位置信息和深层语义信息进行整合, 得到整合后的特征图. 在行人属性识别中, 可以进一步提高小属性识别的准确性.

文献[31]提出了特征金字塔和自下而上的特征采样过程, 又使用横向网络进行目标检测. FPN (feature pyramid networks)包含了自上而下的前馈卷积神经网络方法连接来集成多级特征, 细节如图5所示.

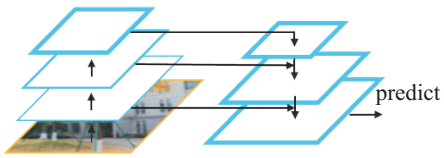


图 5　FPN结构

多尺度特征融合使得目标检测器能够获得更丰富的多尺度特征, 因此, 本文引入多尺度特征金字塔融合策略.

2.4　整体模型

结合金字塔卷积和triplet attention的优点并融合特征金字塔结构, 图6给出了本文所提出的算法.

原始的Resnet 50模型由4个大的残差模块组构成, 每个残差模块组分别由3个、4个、6个和3个残差模块组成, 总共16个残差模块, 每个残差模块由两个1×1卷积和一个3×3卷积组成, 如图7(a)所示. 本文在原始的Resnet 50模型上进行修改, 但主体仍沿用

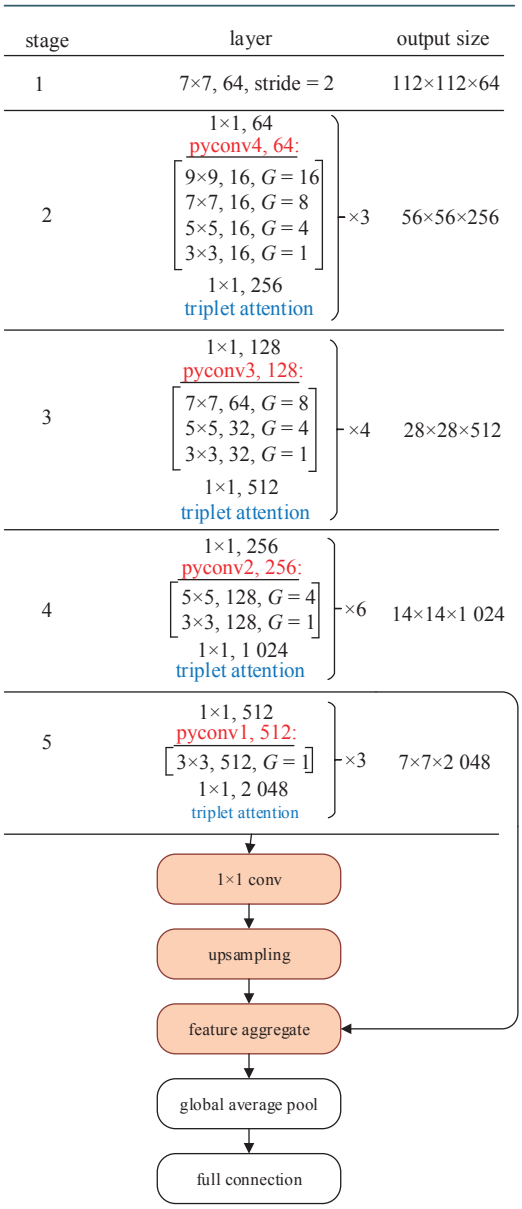
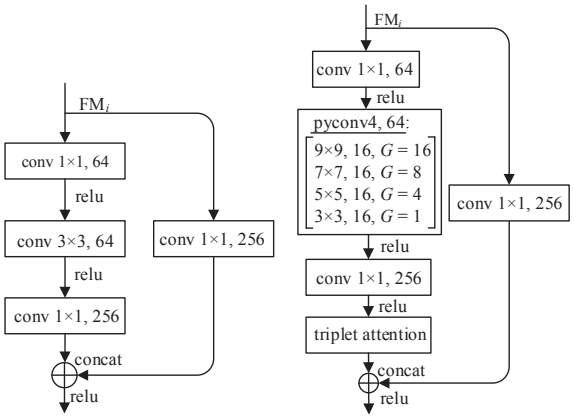


图 6　本文算法



(a) 原始的Resnet 50残差模块　(b) 改进的PyResnet 50残差模块

图 7　残差模块

其架构. 与原始Resnet 50不同, 首先, 在第1级下采样前没有采用最大池化. 然后, 将原始Resnet 50残差模

块中的3×3标准卷积使用金字塔卷积进行替换,在残差模块中插入triplet attention,构成新的残差模块,新的残差模块详细结构如图7(b)所示. 然而,原始的Resnet 50直接采用了第4模块组的特征图. 本文参考特征金字塔的融合策略,将所研究的特征图从第3和第4残差模块组中提取,对特征进行融合的同时提高特征的表示. 最后,获得输出并预测其属性. 与原始的Resnet 50网络相比,本文使用的轻量化多尺度金字塔卷积有效减少了网络计算量和模型大小.

2.5 损失函数

对于多属性分类,通常利用二进制交叉熵损失函数来估计预测值与真实值之间的不一致程度,即

$$L_{\mathrm{CE}}=-\frac{1}{N}\sum_{i=1}^{N}\sum_{l=1}^{L}y_{il}\log(\hat{p}_{il})+(1-y_{il})\log(1-\hat{p}_{il}),$$

$$p_{il}=\frac{1}{1+\exp(-x_{il})}. \tag{6}$$

其中:$\hat{p}_{il}$是样本x_i的第l个属性的估计分数,y_{il}是地面真实标签.

在数据集中,大多数属性类别中正负标签的分布通常是不平衡的. 在训练数据中,许多属性,如太阳镜、塑料袋等,标签相对较少. 直接使用式(6)中的损失函数会由于类的不平衡而损害少属性的预测. 为了减轻这个问题,使用加权交叉熵目标函数,其损失函数为

$$L_{\mathrm{WCE}}=-\frac{1}{N}\sum_{i=1}^{N}\sum_{l=1}^{L}\alpha_l(y_{il}\log(\hat{p}_{il})+(1-y_{il})\log(1-\hat{p}_{il})),$$

$$\alpha_l=\exp\Big(-\frac{p_l}{\gamma^2}\Big). \tag{7}$$

其中:α_l是第l个属性的损失权重,p_l表示训练数据集中第l个属性的正比,γ是一个超参数.

3 实验对比与分析

3.1 数据集与评估方法

3.1.1 数 据 集

本文使用公开数据集PETA和PA100K. PETA共包含19 000幅图像,包括室内和室外场景,分辨率从17×39~169×365像素不等. 实验选取正负比例均衡的35个属性进行训练,19 000幅数据集图像按5∶1∶4比例被随机分为训练集、验证集和测试集. 训练集为9 500幅图像、验证集为1 900幅图像,本实验将训练集和验证集共11 400幅图像结合训练,剩余7 600幅图像用于测试. PA100K数据集通过室外监控摄像头捕获所得,它共包括100 000幅行人图像,分辨率范围为50×100~758×454,并且是目前最大的行人属性识别数据集. 整个数据集按8∶1∶1的比例随机分为训练集、验证集和测试集. 该数据集中的每个图像都用26个属性标记,标签为0或1,分别表示存在或不存在相应的属性.

3.1.2 评估方法

平均精度(mA)是属性识别算法最常用的标准评价. 对于多属性识别中属性的不平衡,将分别计算每个属性的正负样本的识别精度,然后将它们的平均值作为该属性的识别精度,以防止模型偏向那些比例较高的正样本. 具体公式如下:

$$\mathrm{mA}=\frac{1}{2}\sum_{i=1}^{L}\Big(\frac{\mathrm{TP}_i}{P_i}+\frac{\mathrm{TN}_i}{N_i}\Big). \tag{8}$$

其中:L是属性数量,N是样本数量,TP_i和TN_i分别是正确预测的正样本和负样本的数量,P_i和N_i分别是正样本和负样本的数量.

mA是一个基于标签的评估标准,它独立地处理属性,忽略了在多属性识别问题中自然存在的属性间相关性. 因此,文献[32]提出了一种基于实例的评价标准,更符合人类对行人属性预测的一致性. 基于实例的评价标准包括:准确度、精密度、召回率和F_1值,如下所示:

$$\mathrm{accuracy}=\frac{1}{N}\sum_{i=1}^{N}\frac{|Y_i\bigcap f(x_i)|}{|Y_i\bigcup f(x_i)|}, \tag{9}$$

$$\mathrm{precision}=\frac{1}{2N}\sum_{i=1}^{N}\frac{|Y_i\bigcap f(x_i)|}{|f(x_i)|}, \tag{10}$$

$$\mathrm{recall}=\frac{1}{2N}\sum_{i=1}^{N}\frac{|Y_i\bigcap f(x_i)|}{|Y_i|}, \tag{11}$$

$$F_1=\frac{2*\mathrm{prec}*\mathrm{recall}}{\mathrm{prec}+\mathrm{recall}}. \tag{12}$$

其中:N是样本数,Y_i是第i个示例的地面真实正标签,$f(x_i)$是第i个样本的预测正标签,$|\cdot|$表示集合中样本的数量.

3.2 实验设置

本文在Ubuntu16.04、Nvidia GeForce GTX1080Ti等实验条件下,使用CUDA加速,利用PyTorch构建基础网络. 通常在训练模型之前需要对数据进行预处理,这样可以扩展数据集,提高网络的泛化能力. 本文对实验数据的预处理过程包括:将图像大小调整到256×192,以满足网络对输入图像的要求. 训练采用SGD(随机梯度下降),动量为0.9,重量衰减为0.000 5. 初始学习率为0.01,batchsize设置为64,一共训练40个epoch.

为了评估本文方法的有效性和研究各种因素对

行人属性识别的影响,通过使用上述实验设置中的参数,分别在PETA和PA100K数据集上进行多组实验. 本文提出的行人多属性识别网络在PETA数据集上的实验结果如下:基于标签的平均精度为85.81 %,基于实例的准确度为79.75 %,精密度为87.44 %,召回率为86.61 %,F_1值为86.77 %,与其他算法的比较结果如表1所示. 在PA100K数据集上的实验结果如下:基于标签的平均精度为81.48 %,基于实例的准确度为79.92 %,精密度为88.13 %,召回率为87.89 %,F_1值为87.5 %,与其他算法的比较结果如表2所示.

表 1　PETA数据集上不同方法对比结果

method	backbone	PETA				
		mA	accuracy	precision	recall	F_1
ACN[15]	CaffeNet	81.15	73.66	84.06	81.26	82.64
DeepMAR[16]	CaffeNet	82.89	75.07	83.68	83.14	83.41
HPNet[6]	InceptionNet	81.77	76.13	84.92	83.24	84.07
JRL[9]	AlexNet	85.67	–	86.03	85.34	85.42
PGDM[17]	CaffeNet	82.97	78.08	86.86	84.68	85.76
MsVAA[20]	ResNet101	84.59	78.56	86.79	86.12	86.46
MTA-Net[11]	ResNet-152	84.62	78.8	85.67	86.42	86.04
CGCN[13]	ResNet	**87.08**	79.3	83.97	**89.38**	86.59
ours	Resnet 50	85.81	**79.75**	**87.44**	86.61	**86.77**

表 2　PA100K数据集上不同方法对比结果

method	backbone	PA100K				
		mA	accuracy	precision	recall	F_1
DeepMAR[15]	CaffeNet	72.7	70.39	82.24	80.42	81.32
HPNet[6]	InceptionNet	74.21	72.19	82.97	82.09	82.53
PGDM[17]	CaffeNet	74.95	73.08	84.36	82.24	83.29
LGNet[18]	Inception-V2	76.96	75.55	86.99	83.17	85.04
VAC[7]	ResNet50	79.16	79.44	**88.97**	86.26	**87.59**
ALM[19]	BN-Inception	80.68	77.08	84.21	**88.84**	86.46
PD-Net[22]	Inception-V3	80.4	78.8	87.5	86.91	87.2
MCFL[23]	ResNet-50	81.11	79.01	86.67	88.15	87.41
ours	Resnet 50	**81.48**	**79.92**	88.13	87.89	87.5

从表1和表2中可以看出,与现有方法对比,本文提出的模型在RAP和PETA数据集上均取得了不错的结果. 在PETA数据集上的accuracy、precision、F_1,以及在PA100K数据集上的mA、 accuracy等5个指标上实现了最佳性能,每个指标的最高得分均以粗体标出. 在PETA数据集中,ACN[15]和DeepMAR[16]是基于全局的模型,且CaffeNet是早期深度学习模型,提取特征能力较弱. HPNet[6]和MsVAA[20]均使用了注意力提高性能,使得性能提升. JRL[9]和MTA-Net[11]都属于顺序预测方法,PGDM[17]是基于局部的模型. CGCN[13]基于图卷积的方式,它利用特征向量关联属性和图像特征建立属性间的关系. 结果表明,CGCN在mA和recall指标中达到了最优性能,但除recall和mA以外的指标均低于本文方法. 虽然CGCN可以预测足够的属性,但不能保证每个属性的准确性. 原因在于,同一部分可能涉及许多不同的属性,当属性目标较小时,对于同一部分属性的关注度下降,导致相关性的粘合度降低. 而本文提出的方法既挖掘了属性细腻度,又关注属性间的相关性,在总体识别效果上达到了均衡提升. 在PA100K数据集中,LGNet[18]和ALM[19]都属于局部模型,使用了Inception网络. VAC[7]同样基于Resnet 50并使用了注意力. PD-Net[22]旨在获取属性间的依赖关系,MCFL[23]使用不同的损失函数改善属性类间不平衡问题. 结果表明,VAC在Precision和F_1指标上达到了最佳效果,它采用两个分支集成注意机制的方式,有效提升了属性细腻度的识别,但是对于属性间的相关性缺乏考虑,对于每类属性识别效果不佳. ALM[19]自动划分不同区域分别进行预测,提高了各个属性识别准确度,虽然recall指标获得最佳性能,但是本文方法在mA和accuracy指标下都展现了最佳性能,并且在其他指标中都得到了很好的平衡.

3.3　消融实验

本文模型基于Resnet 50结构,具有金字塔卷积、注意模块和特征金字塔集成结构. 为验证本文所提出模型中每个关键组件的有效性以及不同损失函数的影响,使用PETA和PA-100K数据集进行进一步分析,实验对比结果如下. 表3和表4分别给出了各

个模块的有效功能,基于不同的损失函数即CE和WCE进行关键组件的实验对比. 使用原始Resnet 50与不同损失函数作为基线. 两个数据集中的实验表明,在mA指标下,表3中显示,添加特征金字塔融合模块使性能分别提升了1.06 %和1.26 %,表4中提升了0.97 %和1.25 %;在此基础上,表3中添加金字塔卷积模块使性能分别提升了0.84 %和0.34 %,表4中的性能也同样得到了提升. 因为特征融合模块和金字塔卷积都能多尺度处理特征,所以得到的提升是两者相辅的. 然后进行注意力模块的对比,表3中得到了1.06 %和1.03 %的提升,表4中提升了0.2 %和0.86 %,这是因为更加关注了属性区域,且利用属性之间的相关性改善了属性识别. 表3中最终模型性能比基线高了2.2 %和2.7 %,表4中提升了1.55 %和2.41 %. 这些数据均表明了本文所提出的模块中关键组件的出色性能和整体模型的优越性.

表3 使用CE进行的消融实验

dataset	network					mA	accuracy	precision	recall	F_1
	Resnet 50	FA	CE	Pyconv	attention					
PETA	√	×	√	×	×	83.08	74.68	83.16	83.88	83.51
	√	√	√	×	×	84.11	79.39	85.42	84.42	84.91
	√	√	√	√	×	84.95	78.94	86.56	86.43	86.41
	√	√	√	×	√	85.17	79.39	87.14	86.37	86.5
	√	√	√	√	√	85.28	79.59	87.31	86.54	86.67
PA100K	√	×	√	×	×	78.49	75.6	84.86	85.08	84.51
	√	√	√	×	×	79.75	78.78	87.62	86.59	86.72
	√	√	√	√	×	80.09	78.81	87.44	86.85	86.74
	√	√	√	×	√	80.78	78.94	87.08	87.28	86.76
	√	√	√	√	√	81.19	79.25	87.58	87.25	87.04

表4 使用WCE进行的消融实验

dataset	network					mA	accuracy	precision	recall	F_1
	Resnet 50	FA	WCE	Pyconv	attention					
PETA	√	×	√	×	×	84.26	76.09	84.18	84.65	84.12
	√	√	√	×	×	85.23	79.62	87.14	86.53	86.69
	√	√	√	√	×	85.29	79.66	87.35	86.32	86.72
	√	√	√	×	√	85.43	79.69	87.53	86.53	86.76
	√	√	√	√	√	85.81	79.75	87.44	86.61	86.77
PA100K	√	×	√	×	×	79.98	78.05	87.59	86.79	85.24
	√	√	√	×	×	80.32	78.49	86.7	86.96	86.54
	√	√	√	√	×	80.54	79.04	87.7	87.18	86.88
	√	√	√	×	√	81.18	79.38	87.68	87.35	87.13
	√	√	√	√	√	81.48	79.92	88.13	87.89	87.5

3.4 模型大小比较

参考文献[33]的工作,将本文方法与其他方法进行模型大小和性能的实验分析,实验结果如表5所示,其中*表示使用Resnet 50作为骨干网络代替先前原有网络. 在PA100K数据集下,本文方法相比于使用Resnet 50的ALM[19]在模型参数和计算量上分别减少了19.22 %、82.36 %,在mA和F_1指标下分别提升了4.01 %、1.38 %. 本文方法的模型参数和计算量较MsVAA[20]减少了9.5 %和37.74 %,mA和F_1分别提升了1.53 %、2.0 %. 虽然VAC[7]在precision、F_1、指标下比本文高0.84 %和0.09 %,参数量减少了5.29 %,但是mA、accuracy、recall指标分别高了2.32 %、0.48 %

表5 模型性能及大小比较

dataset	methods	backbone	mA	accuracy	precision	recall	F_1	params (M)	MACs (G)
PA100K	ALM[19]*	ResNet50	77.47	75.05	86.61	85.34	85.97	30.86	4.32
	MsVAA[20]*	ResNet50	80.10	76.98	86.26	85.62	85.50	141.27	6.28
	VAC[7]	ResNet50	79.16	79.44	**88.97**	86.26	**87.59**	23.61	14.34
	Ours	ResNet50	**81.48**	**79.92**	88.13	**87.89**	87.50	24.93	3.91

和 1.63 %, 在模型大小相当的情况下, 本文模型计算量减少了 72.26 %, 并且相比于其他算法, 计算量均有明显的下降, 从而表明了本文模型使用金字塔卷积策略的有效性能.

以 GTX1080Ti 作为硬件测试平台, 使用两块 GPU 进行并行加速, 选择 PETA 数据集中的 7 600 张测试数据集进行推理测试. 使用每个图像的平均推理时间, 将时间作为评估算法性能的指标, 取 10 次实验推理时间的平均值作为最终结果, 见表 6. 本文使用原始 Resnet 50 为基线, 使用 WCE 损失函数, 在 PETA 数据集下, 将本文方法与基线方法进行比较, 参数量下降了 2.47 %, 计算量 (浮点运算数) 下降了 5.56 %, 推理时间减少了 34.38 %, mA 提升了 1.55 %; 与使用 Resnet 50 的 ALM[19] 相比, 在模型参数量及计算量上均有减少, mA 和 F_1 指标分别提升了 1.57 %、1.36 %, 在推理时间上减少了 73.78 %. 通过数据显示, 本文方法相比于原始 Resnet 50 和使用 Resnet 50 的 ALM 在推理时间和精度上均有不错的效果.

表 6　模型大小及推理时间比较

methods	PETA					params (M)	MACs (G)	inference time/ms
	mA	accuracy	precision	recall	F_1			
ALM[19]*	84.24	77.84	85.79	85.6	85.41	30.86	4.32	8.81
baseline	84.26	76.09	84.18	84.65	84.12	25.56	4.14	3.52
ours	**85.81**	**79.75**	**87.44**	**86.61**	**86.77**	24.93	3.91	2.31

图 8 为 PETA 数据集中可视化效果. 其中: 横线上半部分表示本文和 DeepMAR[16] 均识别出的结果, 横线下半部分为本文识别出的更多结果. 横线上半部分识别为常见属性且属性明显, 横线下半部分为小属性且具混淆性的属性, 而本文能更准确地识别, 并且对其中鞋子、太阳镜等小目标属性识别效果良好. 由此可见, 本文所提出的模型使用多尺度残差注意的方法提高了属性预测的准确性.

图 8　可视化效果

4　结　论

本文提出的多尺度残差注意的轻量级网络, 有效改善了行人属性的细粒度识别问题, 这是由于金字塔卷积和残差网络出色的特征提取能力和泛化能力, 使算法能够学习多尺度的特征表示, 同时, 通过对特征金字塔整合策略的改进, 使模型能够关注小目标属性. 另外, 使用注意力模块也是一个关键因素, 这使得本文的行人属性识别网络更加关注于图像中存在属性的位置, 推理属性之间的关系, 从而提升整体的识别效果. 本文所提出的网络在预测准确率提升的同时降低了网络参数量和计算量, 实现了准确度和速度的并行优化. 但所做的工作仍存在有待完善的地方, 比如, 将弱监督或者对抗生成网络的方式引进行人属性识别.

参考文献 (References)

[1] Zhu J Q, Liao S C, Lei Z, et al. Pedestrian attribute classification in surveillance: Database and evaluation[C]. IEEE International Conference on Computer Vision Workshops. Sydney, 2013: 331-338.

[2] Lecun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.

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

[4] Chang C C, Lin C J. Libsvm[J]. ACM Transactions on Intelligent Systems and Technology, 2011, 2(3): 1-27.

[5] Lowe D G. Distinctive image features from scale-invariant keypoints[J]. International Journal of Computer Vision, 2004, 60(2): 91-110.

[6] Liu X H, Zhao H Y, Tian M Q, et al. HydraPlus-net: Attentive deep features for pedestrian analysis[C]. IEEE International Conference on Computer Vision. Venice, 2017: 350-359.

[7] Guo H, Zheng K, Fan X C, et al. Visual attention consistency under image transforms for multi-label image classification[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, 2019: 729-739.

[8] Zeng H T, Ai H Z, Zhuang Z J, et al. Multi-task learning via co-attentive sharing for pedestrian attribute recognition[C]. IEEE International Conference on Multimedia and Expo. London, 2020: 1-6.

[9] Wang J Y, Zhu X T, Gong S G, et al. Attribute recognition by joint recurrent learning of context and correlation[C]. IEEE International Conference on Computer Vision. Venice, 2017: 531-540.

[10] Zhao X, Sang L F, Ding G G, et al. Grouping attribute recognition for pedestrian with joint recurrent learning[C]. Proceedings of the 27th International Joint Conference on Artificial Intelligence. Stockholm, 2018: 3177-3183.

[11] Ji Z, Hu Z F, He E L, et al. Pedestrian attribute recognition based on multiple time steps attention[J]. Pattern Recognition Letters, 2020, 138: 170-176.

[12] Li Q Z, Zhao X, He R, et al. Visual-semantic graph reasoning for pedestrian attribute recognition[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2019, 33: 8634-8641.

[13] Fan H N, Hu H M, Liu S L, et al. Correlation graph convolutional network for pedestrian attribute recognition[J]. IEEE Transactions on Multimedia, 2022, 24: 49-60.

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

[15] Sudowe P, Spitzer H, Leibe B. Person attribute recognition with a jointly-trained holistic CNN model[C]. IEEE International Conference on Computer Vision Workshop. Santiago, 2015: 329-337.

[16] Li D W, Chen X T, Huang K Q. Multi-attribute learning for pedestrian attribute recognition in surveillance scenarios[C]. The 3rd IAPR Asian Conference on Pattern Recognition (ACPR). Kuala Lumpur, 2015: 111-115.

[17] Li D W, Chen X T, Zhang Z, et al. Pose guided deep model for pedestrian attribute recognition in surveillance scenarios[C]. IEEE International Conference on Multimedia and Expo. San Diego, 2018: 1-6.

[18] Zhao X, Sang L F, Ding G G, et al. Recurrent attention model for pedestrian attribute recognition[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2019, 33: 9275-9282.

[19] Tang C F, Sheng L, Zhang Z X, et al. Improving pedestrian attribute recognition with weakly-supervised multi-scale attribute-specific localization[C]. IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, 2019: 4996-5005.

[20] Sarafianos N, Xu X, Kakadiaris I A. Deep imbalanced attribute classification using visual attention aggregation[C] Proceedings of the European Conference on Computer Vision (ECCV). Munich, 2018: 680-697.

[21] Zhao X, Sang L F, Ding G G, et al. Recurrent attention model for pedestrian attribute recognition[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2019, 33: 9275-9282.

[22] Liu Y, Tian M Q, Hou J, et al. Pentadent-net: Pedestrian attribute recognition with distance refinement and correlation mining[C]. IEEE International Conference on Image Processing. Abu Dhabi, 2020: 2211-2215.

[23] Zheng X, Yu Z, Chen L, et al. Multi-label contrastive focal loss for pedestrian attribute recognition[C]. The 25th International Conference on Pattern Recognition (ICPR). Milan, 2021: 7349-7356.

[24] Junejo I N. Multi-branch Gabor wavelet layers for pedestrian attribute recognition[J]. IEEE Access, 2021, 9: 40019-40026.

[25] Yu F, Koltun V, Funkhouser T. Dilated residual networks[C]. IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, 2017: 636-644.

[26] Chollet F. Xception: Deep learning with depthwise separable convolutions[C]. IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, 2017: 1800-1807.

[27] Duta I C, Liu L, Zhu F, et al. Pyramidal convolution: Rethinking convolutional neural networks for visual recognition[J/OL]. 2020, arXiv: 2006.11538.

[28] Hu J, Shen L, Albanie S, et al. Squeeze-and-excitation networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 42(8): 2011-2023.

[29] Sanghyun Woo, Jongchan Park, Joon-Young Lee, et al. Cbam: Convolutional block attention module[C]. Proceedings of the European Conference on Computer Vision (ECCV). Munich, 2018: 3-19.

[30] Misra D, Nalamada T, Arasanipalai A U, et al. Rotate to attend: Convolutional triplet attention module[C]. IEEE Winter Conference on Applications of Computer Vision. Waikoloa, 2021: 3138-3147.

[31] Lin T Y, Dollár P, Girshick R, et al. Feature pyramid networks for object detection[C]. IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, 2017: 936-944.

[32] Li D W, Zhang Z, Chen X T, et al. A richly annotated dataset for pedestrian attribute recognition[J/OL]. 2016, arXiv: 1603.07054.

[33] Jia J, Huang H J, Yang W J, et al. Rethinking of pedestrian attribute recognition: Realistic datasets with efficient method[J/OL]. 2020, arXiv: 2005.11909.

作者简介

张再腾 (1997−), 女, 硕士生, 从事计算机视觉、智能图像处理的研究, E-mail: zzteng0466@163.com;

张荣芬 (1977−), 女, 教授, 博士, 从事智能图像处理、人工智能等研究, E-mail: rfzhang@ gzu.edu.cn;

刘宇红 (1963−), 男, 教授, 从事计算机视觉、智能图像处理、大数据智能处理等研究, E-mail: liuyuhongyx@sina.com.

(责任编辑：李君玲)