第 58 卷 第 4 期/2021 年 2 月/激光与光电子学进展 研究论文

激光与光电子学进展 先进成像

基于多尺度注意力网络的行人属性识别算法

李娜[1,2*]，武阳阳[1,2**]，刘颖[2]，邢琎[1]

[1] 西安邮电大学通信与信息工程学院，陕西 西安 710121；

[2] 电子信息现场勘验应用技术公安部重点实验室，陕西 西安 710121

摘要 为了提高行人属性识别的准确率，提出了一种基于多尺度注意力网络的行人属性识别算法。为了提高算法的特征表达能力和属性判别能力，首先，在残差网络 ResNet50 的基础上，增加了自顶向下的特征金字塔和注意力模块，自顶向下的特征金字塔由自底向上提取的视觉特征构建；然后，融合特征金字塔中不同尺度的特征，为每层特征的通道注意力赋予不同的权重。最后，改进了模型损失函数以减弱数据不平衡对属性识别率的影响。在 RAP 和 PA-100K 数据集上的实验结果表明，与现有算法相比，本算法对行人属性识别的平均精度、准确度、F1 性能更好。

关键词 图像处理；行人属性识别；深度学习；特征金字塔；多尺度注意力

中图分类号 TP391.4 **文献标志码** A **doi**：10.3788/LOP202158.0410025

Pedestrian Attribute Recognition Algorithm Based on Multi-Scale Attention Network

Li Na[1,2*], Wu Yangyang[1,2**], Liu Ying[2], Xing Jin[1]

[1] *School of Communication and Information Engineering, Xi'an University of Posts & Telecommunications, Xi'an, Shaanxi* 710121, *China*;

[2] *Key Laboratory of Electronic Information Application Technology for Scene Investigation, Ministry of Public Security, Xi'an, Shaanxi* 710121, *China*

Abstract In order to improve the accuracy of pedestrian attribute recognition, a multi-scale attention network for pedestrian attribute recognition algorithm is proposed in this paper. In order to improve the ability of feature expression and attribute recognition of the algorithm, first, the top-down feature pyramid and attention module are added to the residual network ResNet50. A top-down feature pyramid is constructed from the visual features extracted from the bottom-up. Then, the features of different scales in the feature pyramid are fused to give different weights to the channel attention of each layer of features. Finally, the model loss function is improved to weaken the impact of data imbalance on the attribute recognition rate. Experimental results on the RAP and PA-100K data sets show that compared with existing algorithms, the algorithm has better performance in terms of average accuracy, accuracy, and F1 for pedestrian attribute recognition.

Key words image processing; pedestrian attribute recognition; deep learning; feature pyramid; multi-scale attention

OCIS codes 100.2960; 150.1135; 100.4996; 040.1880

收稿日期：2020-09-27；**修回日期**：2020-10-19；**录用日期**：2020-11-05

基金项目：国家自然科学基金(41874173)、陕西省科技厅双导师制项目(2019JM-604)、西安邮电大学研究生创新基金(CXJJLY2019083)

* **E-mail**：lina114@xupt.edu.cn；** **E-mail**：2745027040@qq.com

1 引　　言

行人属性识别是近年来监控领域的研究热点，目的是识别图像或视频中行人的视觉属性，如年龄、性别、服装风格、鞋子类别，在计算机视觉任务中有很大的应用潜力，如行人重识别[1-2]利用行人属性辅助匹配不同监控摄像头下的同一行人，行人检索[3]利用属性快速检索感兴趣的目标。在许多真实的监控场景中，摄像机被安装在能覆盖一定区域的位置，捕获的行人图像分辨率较低，难以获得清晰的脸部图像。因此，这种情况下的行人属性因其光照不变性和对比度不变性具有很好的应用价值。

行人属性识别算法可大致可分为传统机器学习和深度学习的算法。其中，传统机器学习算法主要由特征提取和分类器设计组成。首先利用手工特征获取行人图像的底层特征，如 Deng 等[4]选取颜色特征、纹理特征以及方向梯度直方图特征，并利用 K 近邻法对特征进行分类。Gray 等[5]通过组合局部手工特征实现行人属性识别。传统机器学习在训练前需要进行手工特征提取，工作量较大，因此不能保证特征选取的合理性。随着深度学习的发展，很多计算机视觉任务[6-8]有了新的突破，Li 等[9]提出的多标签属性学习卷积神经网络（CNN）模型 DeepMar（Deep learning based multiple attributes recognition）也取得了较大的进展。DeepMar 用 CNN 得到更丰富的特征，以代替手工特征，从一个网络模型中同时识别出行人的多个属性。Sudowe 等[10]提出的属性卷积网络（ACN）将整个行人图像作为模型输入，联合学习所有属性的预测。这些方法都是利用网络最后一个特征图完成属性识别任务，不能提高属性识别的准确率，原因是不同网络层的特征反映了属性不同的语义信息。如 CNN 模型需要从前几层网络中提取颜色或纹理等底层特征，这些特征对于衣服颜色和条纹的属性非常重要，但对于性别、年龄这样的语义属性，高层网络的特征比底层特征更有效。Park 等[11]融合多个网络中间层特征识别行人属性，Zhou 等[12]将中间层特征利用不同的内核进行池化操作，以弱监督的方式预测属性的位置，为了有效提高行人属性识别的性能，需要网络中不同层次的特征。

为了应对多层次的属性识别，通过 CNN 提取行人图像的特征构建特征金字塔网络（FPN）[13]。本文针对行人属性识别任务的特点重新设计了多尺度注意力网络，采用基础的残差网络 ResNet50[14]框架构建 FPN，以融合底层特征和高层特征，并利用通道注意力机制提升特征通道之间的关联性，以增强网络的属性识别能力。

2 网络结构

CNN 处理计算机视觉任务时，使用的池化操作会不断缩小特征图的尺寸，网络卷积由底层到高层的分辨率越来越粗糙，特征图也越来越小。但卷积层越高，特征图包含的语义信息越丰富。为了同时利用底层特征和高层特征，本网络的结构如图 1 所示，包含自底向上的特征提取模块、自顶向下构建的 FPN 模块和通道注意力模块，通过融合不同层的特征信息，提高模型对多尺度行人属性的识别效率。

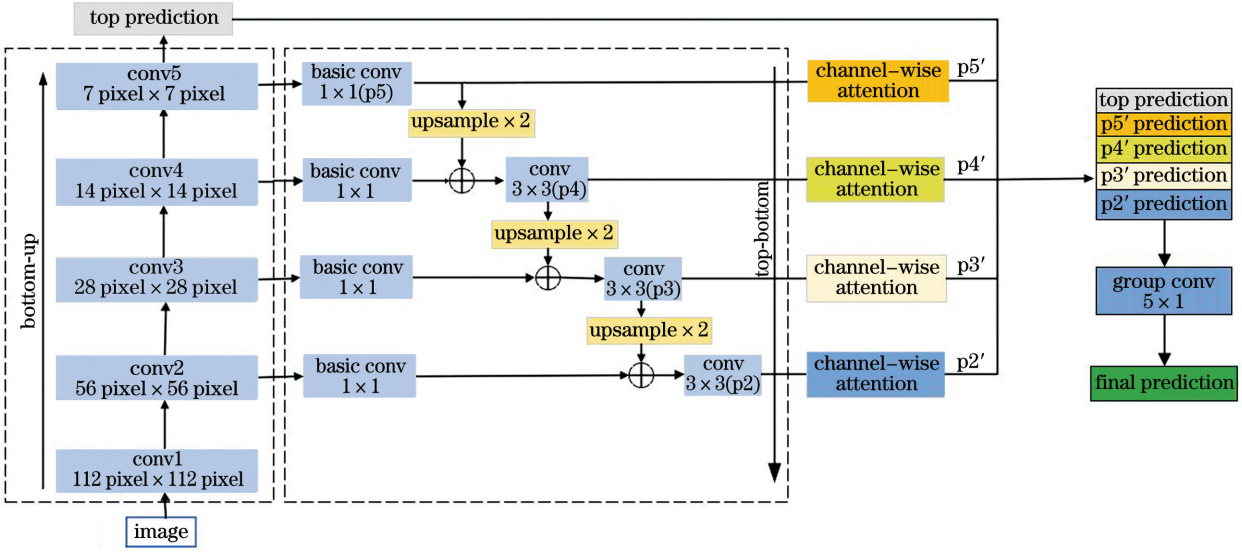

图 1　行人属性识别网络的结构

Fig. 1　Structure of the pedestrian attribute recognition network

2.1　自底向上的特征提取

He 等[14]提出的 ResNet 可以有效解决深度神经网络训练过程中出现的梯度消失以及加深网络导致的退化问题。该网络主要对残差块进行重复使用，将原始输入直接跳过某些层传到之后的层中，并将两者相加作为输出，这样的连接也被称为残差块。实验用 ResNet50 作为主干网络，选择每个阶段卷积层（conv2，conv3，conv4，conv5）最后一个残差结构的特征输出，相对于输入图像的分辨率分别为{56 pixel×56 pixel，28 pixel×28 pixel，14 pixel×14 pixel，7 pixel×7 pixel}。由于 conv1 的特征尺度太大，容易占用内存，因此未被使用。

2.2　自顶向下构建特征金字塔

自顶向下网络的目的是融合相邻层的特征，构建特征金字塔。以 conv3 为例，相邻两层特征的融合过程如图 2 所示，可表示为

$$f_c'=x_{1\times1}(f_c),\tag{1}$$

$$f_u'=X_{\text{upsample}}(f_u)\times2,\tag{2}$$

$$F=x_{3\times3}(f_c'+f_u'),\tag{3}$$

式中，f_c'为经过 basic conv 的横向连接，是一个 1×1 的卷积操作。f_u'为上采样的结果，f_c 为 conv3 层的特征输出，f_u 为 conv4 层与 conv5（相邻）的融合输出，f_c'和 f_u'经像素相加得到融合的特征，再用 3×3 卷积核 $x_{3\times3}$ 进行去卷积处理，得到最终的特征 F。

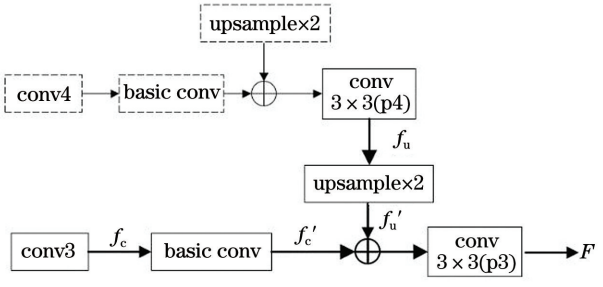

图 2　特征融合的流程图

Fig. 2　Flow chart of the feature fusion

为了保证高层特征图和相邻下一层特征图的大小相同，对高层特征进行 2 倍上采样。实验采用最近邻插值算法减少计算的复杂度。为了在不增加太多计算量的同时融合不同尺度的特征，用 basic conv 改变网络的通道数，由于相邻特征层的通道数不相同，该过程主要包括三个操作：首先将输入特征通过 1×1 卷积层提取；然后经批量正则化（batch norm）层进行数据归一化处理，避免网络提取的特征因数值过大出现不稳定情况；最后经修正线性单元（ReLU）激活函数处理。通过像素间的加法融合相邻层的特征，为了消除上采样带来的混叠效应，用 3×3 的卷积核处理已经融合的特征图。重复迭代该过程，得到 conv2，conv3，conv4，conv5 层对应的融合特征层为 p2，p3，p4，p5。

2.3　通道注意力模块

为了减少网络的参数，利用最大池化对特征图 p5（7×7）进行下采样，如对特征图 p4 用 1 个 3×3 的卷积核进行池化操作，对特征图 p3 则用 2 个 3×3 的卷积核进行池化操作，对特征图 p2 进行类似的池化操作，以保证所有融合的特征层有相同的大小和形状。

由于融合后的最终特征不同，通道的作用也不同，为了增大有效特征通道的权重，减小无效或者效果小的特征通道权重，增强更有利于行人属性识别效果的通道特征表示，结合 Hu 等[15]提出的注意力方法设置一个通道注意力机制，使网络自主学习融合特征中不同的通道特征。最大池化和通道注意力模块结构如图 3 所示，自顶向下的某层特征（如 p5）经过最大池化后得到 x，再利用全局平均池化（GAP）压缩每个通道的特征尺寸。全局平均池化后的特征维度为 $h\times w\times c$，其中，$h\times w$ 为原输入特征图的高度和宽度，c 为属性的数量。该特征经过两个全连接层（FC）的降维和升维，恢复到原来的通道数，以达到编码解码的激励操作，同时极大减少了参数量和计算量。将第二个全连接层的输出经过 Sigmoid 函数得到融合特征中每个通道的注意力权重，最终得到特征 $H(x)$（如 p5′）。

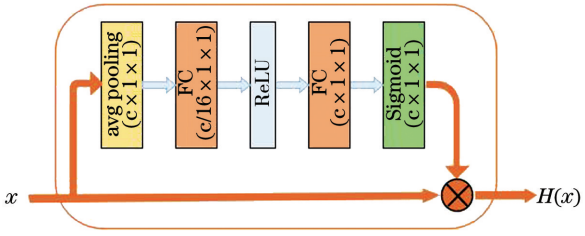

图 3　通道注意力模块

Fig. 3　Channel attention module

最终的预测是 p5′、p4′、p3′、p2′以及自底向上预测（top prediction）的融合结果，实验得到属性的预测为 5 组 N 个属性的特征向量，用 N 个滤波器组、5×1 的卷积核为这 5 组预测属性分配不同的权重，并用 Softmax 函数对每组权重加以约束，确保每组权重的和为 1。最后将注意力加权输出作为属性识别的最终预测结果。

2.4　损失函数

本模型同时识别行人样本的多个属性，本质上是对行人的多标签识别，因此，采用交叉熵损失。在

行人数据集中，如果一张图像具有某个属性，则为该属性的正样本，否则为负样本。但大部分属性的正样本和负样本分布严重不平衡，如 RAP 数据集中，戴眼镜的正样本只占 10%。正样本的比例越小，该属性的损失值越大。为了解决该问题，在损失函数中引入了一个正样本权重因子 w_l。当某个属性的正样本数目较少时，为整个损失函数赋予一个较大的惩罚权重，以防止出现网络训练时因正负样本不平衡导致的梯度爆炸现象，可表示为

$$X_{\text{Loss}}=-\frac{1}{M}\sum_{m=1}^{M}\sum_{l=1}^{L}w_l[y_{m,l}\ln(p_{m,l})+(1-y_{m,l})\ln(1-p_{m,l})], \tag{4}$$

$$p_{m,l}=1/[1+\exp(-x_{m,l})], \tag{5}$$

$$w_l=\exp(-p_l/\sigma^2), \tag{6}$$

式中，M 为行人图像的数量，L 为属性个数，x_m 为不同的行人图像，$m\in1,2,\cdots,M$，对应的属性标签为 y_l，$l\in1,2,\cdots,L$，$y_{m,l}\in\{0,1\}$ 为样本 x_m 的第 l 个属性的标签，$y_{m,l}=1$ 表示样本 x_m 包含第 l 个属性，反之则不包含。$p_{m,l}$ 为属性识别网络对样本 x_m 第 l 个属性预测的结果，p_l 为第 l 个属性的正样本比率，即该属性正样本数目在训练集中所占的比率，σ 为 1。

3　实验结果及分析

3.1　实验数据

实验使用的数据集为 RAP[16] 和 PA-100K[17] 两大行人属性数据集，RAP 数据集在真实室内监视场景拍摄，包含 26 个摄像头拍摄的 41585 张图像，其中，33268 张用于训练，其余图像用于测试，数据集中图像的分辨率从 36 pixel×92 pixel 到 344 pixel×554 pixel。每张图像用 72 个细粒度属性注释，选取正样本比例高于 1% 的 51 个属性进行评价。PA-100K 数据集包括 598 个室外监控摄像头采集到的 100000 张行人图像，分辨率从 50 pixel×100 pixel 到 758 pixel×454 pixel，每张图像用 26 个属性注释。目前 PA-100K 数据集是最大的行人属性识别数据集，将 80000 张图像用于训练，20000 张图像用于测试。

3.2　参数设置和评价指标

实验环境：工作站为 64 位 Ubuntu 系统，服务器的 GPU 为 NVIDIA GTX 1080 Ti、内存为 16 G。模型训练使用的深度学习平台为 Pytorch，编程语言为 Python。根据行人图像特点将输入图像的分辨率统一缩放至 224 pixel×224 pixel，同时使用随机扩张、随机裁剪等数据增强方法，网络使用随机梯度下降法（SGD）训练。共训练 200 个 epoch，batch-size 为 32，学习率（learning-rate）采用多分步策略，初始学习率为 1×10^{-3}，在 150，180 和 200 个 epoch 时学习率依次衰减为上一次的十分之一。参数衰减值（weight-decay）为 0.0005，动量因子（momentum）为 0.9。为防止初始 X_{Loss} 爆炸，前 10 个 epoch 的学习率从 1×10^{-4} 逐渐升至 4×10^{-4}。

衡量行人属性识别能力的两个指标为基于标签的评价指标和基于实例的评价指标。Deng 等[18] 提出基于标签的评价方式即平均精度（mA），分别计算每个属性正样本和负样本识别正确的比例，再将二者的平均值作为每个属性的准确度。Zhao 等[19] 提出基于实例的评价方式，根据每个样本的分对属性和分错属性关系得到 4 个评价指标，分别为准确率（Acc）、精确率（Prec）、召回率（Rec）和 F1。

3.3　网络模型的切片分析

为了更好地分析通道注意力模块的有效性，表 1 对比了基础网络 ResNet50（Baseline）和在每个中间层（conv2、conv3 、conv4 、conv5）添加通道注意力模块的网络（Baseline+CA）对最终属性识别效果的影响，并用 mA 和 F1 评估网络在 RAP 和 PA-100K 数据集上的性能。可以发现，相比 Baseline，Baseline+CA 在 RAP 数据集上的 mA 仅提升了 0.22 个百分点，整体提升效果不明显。原因是监控中行人图像质量差，将通道注意力加入网络中并不能明显关注一些语义特征。

表 1　通道注意力有效性的验证实验

Table 1　Verification experiment of the channel attention effectiveness　　unit：%

Methods	RAP		PA-100K	
	mA	F1	mA	F1
Baseline	75.67	78.20	77.28	84.52
Baseline+CA	75.89	78.36	77.35	84.67

为了验证自顶向下模块的指导效果，在每个数据集上进行 3 组对比实验，第 1 组实验在基础网络中添加图 1 中自顶向下（top-down）的金字塔结构，得到不同阶段的特征（p2，p3，p4，p5），以相加（addition）的方式融合这些特征，相比基础网络（Baseline），简单的相加方式会忽略特征不匹配的问题，包括一些必要的特征，导致属性识别率下降。相比基础网络（Baseline），第 2 组实验在融合不同特征层前赋予不同阶段特征不同权重（weight），在 RAP

和 PA-100K 数据集上的 mA 分别提升了 1.08 和 0.85 个百分点，这表明用高层次特征进行自顶向下的指导是有效果的。第 3 组实验通过引入注意力机制建立特征通道之间的依赖性，在 RAP 和 PA-100K 数据集上的 mA 分别提升了 3.03 和 2.54 个百分点，这说明高层次自顶向下的指导和通道注意力机制的结合有助于提高行人属性的识别率。

表 2 不同特征融合模块的识别结果

Table 2 Recognition results of different feature fusion modules unit: %

Method	RAP		PA-100K	
	mA	F1	mA	F1
Baseline	75.67	78.20	77.28	84.52
Top-down(addition)	74.23	78.01	76.65	84.63
Top-down(weight)	76.75	78.96	78.13	84.95
Top-down(weight)+CA	78.70	80.12	79.82	85.71

3.4 属性在各个特征层的识别率

为了进一步分析网络中多尺度特征融合的有效性，用几个行人属性在 RAP 数据集上进行实验，包括高级语义属性年龄 Age less 16、Age 17～60、Age bigger 60 和低级语义属性上身衬衫(ub-shirt)、下身裙子(lb-skirt)和上身短袖(ub-short sleeve)，结果如表 3 所示。可以看到，如果从自底向上的单个中间层特征(conv2、conv3、conv4、conv5)进行预测，这些高级语义属性以及低级语义属性的 mA 比基线模型(Baseline)差很多，原因是这些特征不足以进行属性预测。而添加自顶向下的金字塔结构及通道注意力模块得到的特征(p5′、p4′、p3′、p2′)可大幅提升识别率。当积累更多不同层次的特征，对 mA 的提升更明显。

对单个属性的分析，推断低级语义属性同时基于低级特征和高级特征。如上身短袖(ub-short sleeve)依赖于从网络的浅层中获得的低层特征，如纹理特征；而上身衣服属性需要从高层特征的语义信息中获得，当高层特征用于指导底层特征时，可大幅度提升属性识别率，这表明高层特征对底层特征具有属性语义指导作用。因此，每个属性的识别需要多层次的特征。由实验数据可知，p5′、p4′、p3′、p2′四个横向预测对不同属性的识别率不同，因此，融合时对所有预测的总和进行加权，结果表明本算法更有效。

表 3 RAP 数据集上每层特征的识别结果

Table 3 Recognition results for each layer feature on RAP data set unit: %

Feature layer	Age less 16	Age 17-60	Age bigger 60	ub-shirt	lb-skirt	ub-short sleeve	mA
conv2	49.54	49.36	50.12	51.23	50.56	54.57	61.29
conv3	50.12	49.78	48.93	50.54	54.18	69.55	65.46
conv4	49.82	49.70	47.65	49.73	53.90	79.78	62.91
p5′	56.76	52.59	60.99	71.93	75.68	77.21	77.01
p4′	61.24	54.42	64.23	78.24	74.76	79.92	77.23
p3′	62.35	56.74	67.42	78.89	77.15	**80.54**	76.89
p2′	62.27	56.67	66.37	77.97	75.43	80.35	77.65
Baseline	63.36	58.49	**71.17**	78.52	78.82	79.15	75.67
Ours	**76.42**	**75.56**	66.92	**79.14**	**80.14**	78.66	**78.70**

3.5 实验结果

本算法与其他行人属性识别算法在 RAP 和 PA-100K 数据集上的训练和验证结果如表 4 和表 5 所示，包括 DeepMar[9]、ACN[10]、联合姿态多属性深度学习(VeSPA)[17]、姿态引导深度模型(PGDM)[20]、基于注意力机制的深度网络(HP-Net)[21]和互动聚集网络(IA2-Net)[22]。其中，DeepMar[9]和 ACN[10]算法都是从一个自底向上的网络中利用最后一层网络信息解决多属性分类问题，缺乏对属性中先验信息的考虑。IA2-Net[22]采用卷积神经网络-循环神经网络(CNN-RNN)模型作为主要的体系结构，用行人图像和相应的属性作为输入，并设计了一种融合注意力机制，使模型能对输入进行加权。VeSPA[17]通过粗略的视图预测器进行属性预测。PGDM[20]使用预先训练的姿态估计模型估计图像中人体的关键点，最后将局部关键点提取的特征与全局特征融合起来进行属性识别。这类方法需要额外的网络信息，不能进行端到端的训练，且模型的复杂度较高。HP-Net[21]利用多方向注意力模块将注意力机制应用于多层图像特征，

也可以获得行人属性识别从低级到高层语义的有效信息。本算法使用高层次语义信息指导底层次信息，并利用通道注意力机制建立属性之间的依赖关系。

表 4 不同算法在 RAP 数据集上的识别结果

Table 4 Recognition results of different algorithms on the RAP data set unit: %

Algorithm	mA	Acc	Prec	Rec	F1
ACN	69.66	62.61	80.12	72.26	75.98
DeepMar	73.79	62.02	74.92	76.21	75.56
HP-Net	76.12	65.39	77.33	78.79	78.05
VeSPA	77.70	67.35	79.51	79.67	79.59
PGDM	74.31	64.57	78.86	75.90	77.35
IA2-Net	77.44	65.75	**79.01**	77.45	78.03
Ours	**78.70**	**68.17**	78.89	**79.98**	**80.12**

表 5 不同算法在 PA-100K 数据集上的识别结果

Table 5 Recognition results of different algorithms on the PA-100K data set unit: %

Algorithm	mA	Acc	Prec	Rec	F1
DeepMar	72.70	70.39	82.24	80.42	81.32
HP-Net	74.21	72.19	82.97	82.09	82.53
VeSPA	76.32	73.00	84.99	81.49	83.20
PGDM	74.95	73.08	84.36	82.24	83.29
IA2-Net	77.28	74.73	83.34	**85.73**	84.52
Ours	**79.82**	**78.17**	82.83	84.98	**85.71**

从表 4 和表 5 可以发现，与现有基于标签和基于实例的度量算法相比，本算法在两个数据集上都取得了良好的性能，这也验证了多尺度融合的有效性。相比 IA2-Net 算法，在更具有挑战性的数据集 PA-100K 上，本算法的 mA 和 F1 提高了 2.54 和 1.19 个百分点；在 RAP 数据集中的 mA 和 F1 提高了 1.26 和 2.09 个百分点。需要注意的是，本算法在召回率较高时精确率较低，原因是这两个度量是负相关，即一个度量的增加总是导致另一个度量的减少，特别是调节损失函数权重时。因此这两个指标对于样本不平衡时的评估结果不是很可靠，在评价属性识别模型的性能时，mA 和 F1 评价指标更合适，而本算法在这两个指标中的结果是最好的。

4 结 论

提出了一种端到端多尺度特征融合的行人属性识别算法，通过对残差网络建立特征金字塔以及通道注意力模块，有效融合了高层语义特征和底层特征，进一步提升了模型的学习效率。在 RAP 和 PA-100K 数据集上的实验结果表明，与现有算法相比，本算法的平均精度、准确度和 F1 性能更好。但该模型没有充分利用属性之间的联系，后续可将属性之间的关系建模整合到行人属性识别网络中，进一步提高行人属性的识别率。

参 考 文 献

[1] Schumann A, Stiefelhagen R. Person re-identification by deep learning attribute-complementary information [C]//2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 1435-1443.

[2] Su C, Yang F, Zhang S L, et al. Multi-task learning with low rank attribute embedding for multi-camera person re-identification [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(5): 1167-1181.

[3] Schumann A, Specker A, Beyerer J. Attribute-based person retrieval and search in video sequences [C]// 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), November 27-30, 2018, Auckland, New Zealand. New York: IEEE Press, 2018: 1-6.

[4] Deng Y B, Luo P, Loy C C, et al. Pedestrian attribute recognition at far distance [C]//Proceedings of the 22nd ACM International Conference on Multimedia, November 3-7, 2014, Orlando, Florida, USA. New York: ACM, 2014: 789-792.

[5] Gray D, Tao H. Viewpoint invariant pedestrian recognition with an ensemble of localized features [M]// Forsyth D, Torr P, Zisserman A, et al. Computer Vision-ECCV 2018. Lecture Notes in Computer Science. Cham: Springer, 2018, 5302: 262-275.

[6] Chen L L, Zhang Z D, Peng L. Real-time detection based on improved single shot MultiBox detector [J]. Laser & Optoelectronics Progress, 2019, 56(1): 011002.
陈立里, 张正道, 彭力. 基于改进 SSD 的实时检测方法[J]. 激光与光电子学进展, 2019, 56(1): 011002.

[7] Wang J Q, Li J S, Zhou X W, et al. Improved SSD algorithm and its performance analysis of small target detection in remote sensing images [J]. Acta Optica Sinica, 2019, 39(6): 0628005.
王俊强, 李建胜, 周学文, 等. 改进的 SSD 算法及其

对遥感影像小目标检测性能的分析[J]. 光学学报, 2019, 39(6): 0628005.

[8] Ou P, Zhang Z, Lu K, et al. Object detectionin of remote sensing images based on convolutional neural networks[J]. Laser & Optoelectronics Progress, 2019, 56(5): 051002.
欧攀, 张正, 路奎, 等. 基于卷积神经网络的遥感图像目标检测[J]. 激光与光电子学进展, 2019, 56(5): 051002.

[9] Li D W, Chen X T, Huang K Q. Multi-attribute learning for pedestrian attribute recognition in surveillance scenarios[C]// 2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR), November 3-6, 2015, Kuala Lumpur, Malaysia. New York: IEEE Press, 2015: 111-115.

[10] Sudowe P, Spitzer H, Leibe B. Person attribute recognition with a jointly-trained holistic CNN model[C]// 2015 IEEE International Conference on Computer Vision Workshop (ICCVW), December 7-13, 2015, Santiago, Chile. New York: IEEE Press, 2015: 329-337.

[11] Park S, Zhu S C. Attributed grammars for joint estimation of human attributes, part and pose[C]// 2015 IEEE International Conference on Computer Vision (ICCV), December 7-13, 2015, Santiago, Chile. New York: IEEE Press, 2015: 2372-2380.

[12] Zhou Y, Yu K, Leng B, et al. Weakly-supervised learning of mid-level features for pedestrian attribute recognition and localization[EB/OL]. [2020-09-15]. http://arxiv.org/abs/1611.05603.

[13] 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 (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 936-944.

[14] He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition[C]// 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE Press, 2016: 770-778.

[15] Hu J, Shen L, Albanie S, et al. Squeeze-and-excitation networks[EB/OL]. [2020-09-11]. https://arxiv.org/abs/1709.01507.

[16] Li D W, Zhang Z, Chen X T, et al. A richly annotated dataset for pedestrian attribute recognition[EB/OL]. [2020-09-15]. https://arxiv.org/abs/1603.07054.

[17] Saquib M S, Schumann A, Wang Y, et al. Deep view-sensitive pedestrian attribute inference in an end-to-end model[EB/OL]. [2020-09-13]. http://arxiv.org/abs/1707.06089.

[18] Deng Y B, Luo P, Loy C C, et al. Learning to recognize pedestrian attribute[EB/OL]. [2020-09-18]. http://arxiv.org/abs/1501.00901.

[19] Zhao X, Sang L F, Ding G G, et al. Grouping attribute recognition for pedestrian with joint recurrent learning[C]// Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, July 13-19, 2018, Stockholm, Sweden. California: AAAI Press, 2018: 3177-3183.

[20] Li D W, Chen X T, Zhang Z, et al. Pose guided deep model for pedestrian attribute recognition in surveillance scenarios[C]// 2018 IEEE International Conference on Multimedia and Expo (ICME), July 23-27, 2018, San Diego, CA, USA. New York: IEEE Press, 2018: 1-6.

[21] Liu X H, Zhao H Y, Tian M Q, et al. HydraPlus-net: attentive deep features for pedestrian analysis[C]// 2017 IEEE International Conference on Computer Vision (ICCV), October 22-29, 2017, Venice, Italy. New York: IEEE Press, 2017: 350-359.

[22] Ji Z, He E L, Wang H R, et al. Image-attribute reciprocally guided attention network for pedestrian attribute recognition[J]. Pattern Recognition Letters, 2019, 120: 89-95.