面向行人重识别的局部特征研究进展、挑战与展望

姚 足[1] 龚 勋[1] 陈 锐[1] 卢 奇[1] 罗 彬[1]

摘 要 行人重识别 (Person re-identification, Re-ID) 旨在跨区域、跨场景的视频中实现行人的检索及跟踪, 其成果在智能监控、刑事侦查、反恐防暴等领域具有广阔的应用前景. 由于真实场景下的行人图像存在光照差异大、拍摄视角不统一、物体遮挡等问题, 导致从图像整体提取的全局特征易受无关因素的干扰, 识别精度不高. 基于局部特征的方法通过挖掘行人姿态、人体部位、视角特征等关键信息, 可加强模型对人体关键区域的学习, 降低无关因素的干扰, 从而克服全局特征的缺陷, 也因此成为近几年的研究热点. 本文对近年基于局部特征的行人重识别文献进行梳理, 简述了行人重识别的发展历程, 将基于局部特征的方法归纳为基于姿势提取、基于特征空间分割、基于视角信息、基于注意力机制四类, 并详细阐述了每一类的原理及优缺点. 然后在三个主流行人数据集上对典型方法的识别性能进行了分析比较, 最后总结了目前基于局部特征算法的难点, 并对未来本领域的研究趋势和发展方向进行展望.

关键词 行人重识别, 局部特征, 深度学习, 计算机视觉

引用格式 姚足, 龚勋, 陈锐, 卢奇, 罗彬. 面向行人重识别的局部特征研究进展、挑战与展望. 自动化学报, 2021, **47**(12): 2742−2760

DOI 10.16383/j.aas.c190821

开放科学(资源服务)标识码(OSID):

Research Progress, Challenge and Prospect of Local Features for Person Re-Identification

YAO Zu[1] GONG Xun[1] CHEN Rui[1] LU Qi[1] LUO Bin[1]

Abstract Person re-identification (Re-ID) aims to achieve pedestrian retrieval and tracking in cross-region and cross-scene video. Its achievements have broad application prospects in intelligent monitoring, criminal investigation, counter-terrorism and riot control. Due to pedestrian images in real scenes having problems such as large illumination differences, different shooting angles, and object occlusion, the global feature is susceptible to interference from irrelevant factors, resulting in low recognition accuracy. The local feature-based method strengthens the model′s learning of key areas of the human body and reduces the interference of irrelevant factors by mining key information such as pedestrian posture, human body parts, and perspective features. Because the local feature method overcomes the defect of the global feature, it has become a research focus in recent years. In this paper, we combed the literature of Re-ID based on local features in recent years, and briefly described the development process of Re-ID. The methods based on local features can be classified into four categories: postural extraction, feature spatial partition, viewpoint information and attention mechanism. This paper first elaborates on the principles, advantages and disadvantages of each category. Then we summarize some typical methods in detail and compare their performance on three mainstream Re-ID data sets. Finally, this paper summarizes the difficulties of the method based on local features, and looks forward to the future research trend and development direction of this field.

Key words Person re-identification (Re-ID), local feature, deep learning, computer vision

Citation Yao Zu, Gong Xun, Chen Rui, Lu Qi, Luo Bin. Research progress, challenge and prospect of local features for person re-identification. *Acta Automatica Sinica*, 2021, **47**(12): 2742−2760

行人重识别 (Person re-identification, Re-ID), 也称行人再识别, 旨在利用计算机技术将同一个人在不同地点、不同摄像头捕获的图像关联起来, 从而实现跨监控图像、设备视频中的行人检索及轨迹跟踪. 行人重识别技术克服了人工检索的低效性, 弥补了目前固定摄像头的视觉局限性, 具有巨大的实用价值和应用前景. 目前的行人重识别有着广泛的实际应用, 其中具有代表性的有智能安防领域的嫌犯追踪、大型公共场所的智能寻人、智慧商业的无人超市、智能机器人领域等. 同时, 行人重识别可

收稿日期 2019-12-03 录用日期 2020-04-27
Manuscript received December 3, 2019; accepted April 27, 2020
国家自然科学基金 (61876158), 四川省重点研发项目 (2019YFS0432) 资助
Supported by National Natural Science Foundation of China (61876158) and Sichuan Science and Technology Program (2019YFS0432)
本文责任编委 刘青山
Recommended by Associate Editor LIU Qing-Shan
1. 西南交通大学计算机与人工智能学院 成都 611756
1. School of Computing and Artificial Intelligence, Southwest Jiaotong University, Chengdu 611756

与人脸识别、步态识别、语义分析、属性识别等其他领域的技术结合, 应用于目标行人的跨地域跟踪、自然语言行人检索等任务.

行人重识别在实际应用中面临诸多难点, 主要表现在姿势、步态、服装等行人属性多变, 及光照变化、摄像视角差异、物体遮挡等环境因素干扰严重. 这些差异性导致提取人体的鲁棒特征表示极为困难, 其中影响行人重识别性能的最大因素是行人姿态变化及物体遮挡. 图 1 为不同的监控角度下及遮挡场景下的行人图像, 可以观察到由于不同视角以及物体遮挡的干扰, 摄像头不能捕获到完整的行人图像, 这使得 PAN (Pedestrian alignment network)[1]、Transfer[2]、SOMAne (Somatotype network)[3] 等依赖图像整体信息的全局特征方法失效.

(a) 不同行人的物体遮挡情况
(a) The occlusion of different pedestrians

(b) 同一行人在不同视角下的差异
(b) Differences of the same pedestrian in different view points

图 1　不同视角下及遮挡场景下的行人图像
Fig. 1　Pedestrian images in different viewpoints and occlusion scenes

为了解决全局特征的缺陷, 提取具有更能表征细节信息的局部特征成为研究的热点. 常见的局部特征有骨架、姿势、人体部件等, 这些关键区域的特征可以辅助模型更加精准的区分行人特征与无关特征. 通过综合分析, 本文将基于局部特征的方法归纳为 4 类:

1) 结合行人姿势. 通过额外的人体姿势或骨架预测模型提取人体关键点, 然后将关键点特征与行人重识别模型融合, 生成精确的人体语义部件 (头、身、手、脚等) 区域, 最后针对关键区域的特征匹配.

2) 特征空间分割. 常用的分割方式包括网格分割和水平分割, 将特征图均匀划分得到一系列显著性区域, 使模型对每一个区域的单独训练, 学习人体不同区域的差异.

3) 整合视角信息. 不同角度观测到的人体存在较大的姿态偏差, 如俯视、侧视等角度下的行人外观有较大偏差. 反过来利用视角信息, 在不同角度下建模可使行人重识别方法适应更复杂的拍摄场景.

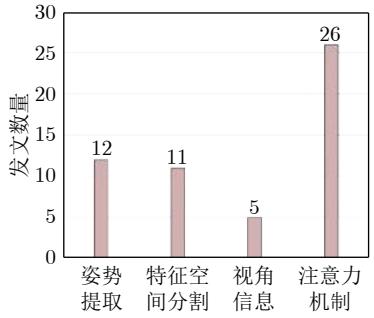
4) 融合注意力机制. 注意力机制能够指导模型重点关注图像的特定区域, 合理融合原始特征与注意力模块可促进模型自主学习关键区域.

按照输入数据的类型, 行人重识别的研究工作可分为单帧图像与视频序列的方法, 考虑到目前主流研究以单帧图像方法为主, 且基于视频序列的方法在设计思路、评估实验及实验性能都与单帧图像方法有较大差距, 因此本文主要整理了单帧图像方法的相关文献. 单帧图像方法可分为有监督学习方法及无监督、半监督学习方法, 本文将有监督学习方法归纳为基于全局特征的方法和基于局部特征的方法, 并在第 4 节中分析比较以上各类方法的性能表现. 在文献调研方面, 本文采用文献法选用 82 篇行人重识别相关领域的高引用文献, 主要的来源为 *IEEE Transactions* (*Institute of Electrical and Electronics Engineers Transactions*)、*PAMI* (*Pattern Analysis and Machine Intelligence*)、*IJCV* (*International Journal of Computer Vision*) 等期刊, CVPR (Conference on Computer Vision and Pattern Recognition)、ECCV (European Conference on Computer Vision)、ICCV (International Conference on Computer Vision) 等计算机视觉顶会. 为了全面把握局部特征在行人重识别领域的发展趋势, 本文统计了 2016 年~2019 年计算机视觉顶会 CVPR、ECCV、ICCV中行人重识别文章总计 159 篇, 其中基于局部特征的行人重识别方法共 54 篇, 图 2 按照局部特征的分类进行梳理, 本文将对这四类基于局部特征的方法进行综述分析.

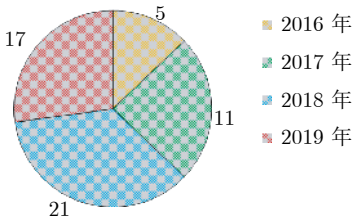
本文的组织结构安排如下: 第 1 节概述了行人重识别工作流程并回顾了行人重识别关键技术的发展历程. 第 2 节介绍了行人重识别领域的常用数据集, 同时简述了行人重识别领域的主要评估方法. 第 3 节详细分析了 4 类基于局部特征的行人重识别方法. 第 4 节针对现有的各类基于局部特征的行人重识别方法, 在公开数据集上分析比较了各类算法中性能优秀的代表模型. 第 5 节探讨了行人重识别研究当前所面临的挑战, 并展望未来值得关注的方向.

1　背景及相关工作

行人重识别的根本问题是确定目标行人与被检测行人的相似度. 传统的行人重识别任务如图 3 所示, 其核心技术主要包括特征的提取与表达, 特征相似性度量. 首先选择适当的特征模型提取的表达能力强的行人特征, 获得特征表示向量, 如 ELF



(a) 各类方法的文献数量
(a) The literature statistics of each method



(b) 局部特征方法在各年份的文献数量
(b) The literature statistics of local feature in each year

图 2 4 类基于局部特征的行人重识别方法文献统计
Fig. 2 Literature statistics of four kinds of local feature-based Re-ID methods

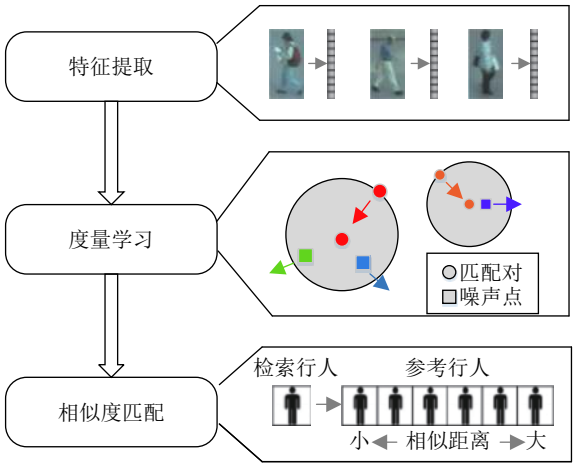


图 3 传统的行人重识别任务执行流程
Fig. 3 The pipline of traditional Re-ID task

(Ensemble of localized features)[4]、SDALF (Symmetry-driven accumulation of local features)[5−6]、SIFT (Scale invariant feature transform)[7]、HOG (Histogram of oriented gradient)[8]、LBP (Local binary pattern)[9] 等颜色或纹理特征; 然后借助 KISSME (Keep it simple and straightforward metric learning)[10]、XQDA (Cross-view quadratic discriminant analysis)[11] 等度量学习方法约束特征间的距离关系, 使同一类别特征间的类内距离尽可能小, 不同类别特征间的类间距离尽可能大; 最后筛选检索目标的近邻特征作为匹配结果.

传统方法将上述的特征提取和相似性度量作为两个互相独立的任务, 因此早期的行人重识别研究主要集中在如何设计表达能力好的手工特征、鲁棒性强的度量学习算法上. 而深度学习方法将行人重识别任务统一到一个网络框架中, 实现端到端的处理流程. 可以概括为:

1) 设计合理的卷积神经网络, 通过前向卷积运算提取特征.

2) 确定优化网络的目标函数, 通过反向传播传播算法更新网络中的参数, 约束模型对特征的学习.

3) 计算特征间的相似度, 按照相似程度排序候选行人.

按照时间点划分, 行人重识别技术的发展可以分为两个阶段: 2014 年前的手工特征阶段和 2014 年后的深度学习阶段. 随着计算机视觉领域的不断发展, 行人重识别技术经历了从通用特征向专门特征, 从手工特征向深度特征的不断演化, 并取得了一系列里程碑性的突破, 部分关键技术发展过程见图 4.

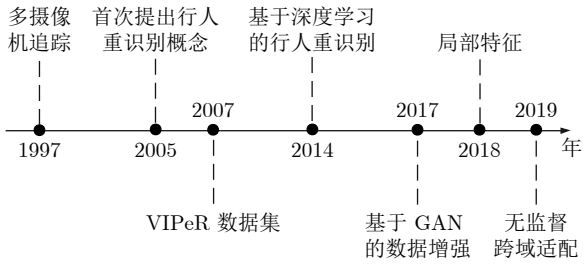


图 4 行人重识别发展中的关键技术
Fig. 4 Key technologies in the development of Re-ID

1) 1997 年, Huang 等[12] 提出了跨摄像头行人追踪, 利用贝叶斯概率公式预测摄像头中出现行人的外观属性, 将行人重识别思想首次应用于学术研究.

2) 2005 年, Zajdel 等[13] 在多摄像头跟踪工作中提出 “Person re-identification”的概念. 这项工作标志着行人再识别领域与多摄像头跟踪领域的研究分离, 在此后行人重识别成为一个独立的计算机视觉任务.

3) 2007 年, Gray 等[14] 提出了 VIPeR 数据集, 更全面地考虑了真实场景中的视角、光线等因素, 吸引了更多的学者投身行人重识别的研究, 为早期的深度学习研究奠定基础.

4) 2014 年, Li 等[15] 卷积神经网络构建 FPNN (Filter pairing neural network) 模型, 能够在统一的框架下完成特征提取、特征匹配任务, 实现了端

到端的行人重识别. 并且结合交叉熵损失 (Cross entropy loss), 将行人重识别引入到深度学习领域.

5) 2017 年, Zheng[16] 开始使用生成对抗网络 (Generative adversarial network, GAN) 改进现有数据集, 通过迁移学习扩增海量数据. 行人重识别的数据获取困难, 这项研究为数据获取提供了全新的思路.

6) 2018 年, Sun 等[17] 提出了基于图像水平分割的局部特征算法, 更多的工作再此基础上继续改进, 使得行人重识别模型的准确率取得了飞跃性的提升, 切实地证明了局部特征在行人重识别模型中的重要价值.

7) 2019 年, 无监督学习的研究取得了显著的进展, 大量研究[18−20] 开始使用无监督方法解决行人重识别任务中的跨域问题. 由于不需要额外标注信息, 无监督学习兼具高效性和便捷性, 对工业界应用有很高的研究意义.

得益于计算机视觉领域的快速发展, 行人重识别与更多研究领域交融, 每年都会集中涌现大量针对现存某一难题的研究. 例如生成对抗网络解决数据稀缺问题、无监督学习处理数据标注问题等. 总的来说, 目前的行人重识别正处于逐步发展成熟的阶段, 呈现多模态、多领域结合的研究趋势.

2 行人重识别公共数据集及评估方法

2.1 常用数据集简介

公开数据集的发展在一定程度上反映了研究领域的发展, 规模更大、场景更复杂的数据集使行人重识别技术得以快速发展. 目前的主流行人重识别数据库如下:

1) VIPeR, 它是早期具有代表性的小型数据集, 包含 632 个行人, 1 264 张行人图像. 由于早期的行人重识别缺乏数据支持, 该数据集的发布引起了更多研究者关注行人重识别领域, 同时为深度学习的发展奠定了基础.

2) Market-1501[21], 包含 1 501 个行人, 33 217 幅行人图像, 由清华大学校园内的 6 个摄像头拍摄得到. 每个行人都至少被 2 个摄像头捕获到, 行人不同视角以及姿势数据十分丰富, 其训练集包含 751 个行人, 平均每个行人有 17 幅图像, 测试集包含 750 个行人, 平均每个行人有 26 幅图像.

3) CUHK03[15], 采集于香港中文大学校园, 该数据集包含来自 6 个摄像机的 1 467 人的 14 097 幅图像. 该数据集中提供了两种类型的标注: 手动标记行人边界框和 DPM 检测到的边界框. 最后将整个数据集随机分成 20 个部分进行交叉验证, 对深度学习方法进行测试时会消耗大量时间.

4) DukeMTMC-reID, 它是 DukeMTMC[22] 的子集, 由 8 台高分辨率相机拍摄的 36 411 幅 1 812 行人的图像组成. 从数据集中随机抽取 702 人的 16 522 幅图像作为训练集, 将剩余的 702 人划分为包含 2 228 张查询图像和 17 661 幅图像的测试集. 数据集中不同行人之间的相似度高, 而同一行人的多幅图像外观差异较大, 使得性能测试较为严苛.

5) MSMT17[23], 由北京大学在 2018 年 CVPR 上提出, 图像采集至 12 个户内摄像头和 3 个户外摄像头. 该数据集是目前最大的单帧行人重识别数据集, 包含 4 101 个行人, 总计 126 441 幅图像. 并且数据集的图像时间跨度大, 涵盖了 4 种不同天气下的早上、中午、下午四个时间段, 更贴近真实场景.

以上列举的数据集都属于单帧行人数据集, 是目前学术界研究的主要对象. 其中 VIPeR 由于采集较早, 其图像分辨率较低, 导致识别难度较大; MSMT17 是目前最大的单帧行人数据集, 其图像质量贴近真实场景, 同时也缺少严格的数据清洗, 使目前的方法尚难取得良好的识别效果, 具有较大的挑战性; Market-1501、CUHK03、DukeMTMC-ReID 均采集至大学校园, 图像数量较多且质量较高, 因此目前的大量工作都基于这三个数据集进行实验. 除以上单帧数据集, 还有针对真实场景的视频数据集 MARS[24]、iLIDS-VID[25]、PRID2011[26] 等. 以及针对遮挡研究常用的小型数据集 Partial-REID[27] 和 Partial-iLIDS[28]. 以上数据集的详细信息可以在表 1 中查阅.

现存的行人重识别数据集虽然系统的扩充了各种类型的优质数据, 在一定程度上解决了早期缺乏实验数据的问题, 但仍然存在不足:

1) 数据量问题. 相较于人脸识别, Re-ID 数据集采集难度更加困难, 因此目前主流数据集中的样本容量也较少. 过少的样本容量导致现有的许多深度学习方法出现严重的过拟合现象, 无法有效地判断方法的正确性.

2) 遮挡问题. Re-ID 中遮挡问题一直难以解决, 一个很大的原因是缺少针对性的数据集进行实验, 无法进一步验证方法的有效性. 目前的遮挡数据集如 Partial-REID 和 Partial-iLIDS 等只有数百张图像, 研究欠缺稳定的分析手段.

3) 视角问题. 在真实场景下的拍摄情况更加复杂, 需要考虑俯视、斜视等各个角度的情况. 而目前公开数据集中的行人大部分在平行视角下采集, 欠缺立体视角的数据样本.

4) 光照变化问题. 光照变化会造成图像分辨率

表 1 行人重识别主流数据集
Table 1 Mainstream Re-ID dataset

库名	发布机构	样本描述	类型	示例
VIPeR (2008)	加州大学圣克鲁兹分校	632 个行人, 1 264 幅行人图像	单帧数据集	
PRID2011 (2011)	格拉茨技术大学	934 个行人, 24 541 帧行人图像,	视频数据集	
Partial-iLIDS (2011)	伦敦玛丽女王大学	119 个行人, 238 幅行人图像	单帧遮挡数据集	
iLIDS-VID (2014)	伦敦玛丽女王大学	300 个行人, 42 495 帧行人图像	视频数据集	
Duke MTMC-reID (2014)	杜克大学	1 812 个行人, 36 441 幅行人图像	单帧数据集	
Partial-ReID (2015)	中山大学	60 个行人, 600 帧行人图像,	单帧遮挡数据集	
Market-1501 (2015)	清华大学	1 501 个行人, 33 217 幅行人图像	单帧数据集	
MARS (2016)	悉尼大学	1 261 个行人, 1 191 003 帧行人图像	视频数据集	
CHUK03 (2017)	香港中文大学	1 467 个行人, 13 164 幅行人图像	单帧数据集	
MSMT17 (2018)	北京大学	4 101 个行人, 126 441 幅行人图像	单帧数据集	

及行人服装出现较大差异, 增大识别难度. 而现存的公开数据集都是在街景下随机采集, 难以获取一个人在不同时段、不同天气下的数据, 因此数据集中的光线差异较小.

2.2 评估方法

目前主流的 Re-ID 数据集主要使用两种性能评估标准. 第 1 种是平均准确度 (Mean average precision, mAP). mAP 首先建立在准确率 (Percision rate) 和召回率 (Recal rate) 之上, 定义为

$$PR = \frac{TP}{TP + FP} \tag{1}$$

$$RR = \frac{TP}{TP + FN} \tag{2}$$

其中, TP, FP, FN 分别表示系统正确判断的正样例、系统错误判断的正样例和系统错误判断的负样例. PR 和 RR 的数值一般呈现相互制约的趋势, 准确度会随着召回率的增大而下降, 单纯使用其中的某一个值都不能全面的评价系统. 一个理想的系统应当在准确度较高的同时, 召回率也尽可能的高. 为了达到这个目的, mAP 使用 P-R 曲线与坐标轴之间的面积来表示平均准确度, 反映了全局信息的情况, 定义为

$$mAP = \int_0^1 P(R)\mathrm{d}R \tag{3}$$

第 2 种是累计匹配性能 (Cumulative match characteristics, CMC) 曲线. 对于查询集 $Q = \{q_1, q_2, \cdots, q_M\}$ 和候选集 $G = \{g_1, g_2, \cdots, g_N\}$, 行人重识别的任务是找到特征相似度较高的 q_i 和 g_j. 将匹配结果按照相似度排序, 然后使用 CMC 曲线的计算模型的命中率.

$$CMC(K) = \frac{1}{N}\sum_{i=1}^{N}\begin{cases} 1, & k_i \le K \\ 0, & k_i > K \end{cases} \tag{4}$$

式中, k_i 表示第 i 个行人的第 k 个匹配结果, CMC 曲线反映了前 k 个匹配结果正确的概率. 当 $K = 1$ 时, 用 rank-1 表示首位命中率, 常用于评价行人重识别算法的优劣性. 而在实际场景中也可以选择 rank-5 和 rank-10 筛选多幅图片, 结合人工判别以提高命中率.

3 局部特征的行人重识别模型分类综述

行人重识别可以分为基于全局特征的方法和基于局部特征的方法. 全局特征只考虑了图像的整体信息, 易受外界环境等无关信息的影响, 很难适应复杂场景下的任务. 局部特征模型更关注骨架、姿势、人体部件等关键区域, 具有更好的抗干扰能力. 本文依据行人局部特征的类别和特点, 将基于局部特征的行人重识别方法分为四类: 基于姿势提取、基于特征空间分割、基于视角信息、基于注意力机制.

3.1　基于姿势估计的方法

3.1.1　姿势估计模型

受拍摄角度不同和行人运动的影响, 行人不同时刻的姿态差异较大, 导致匹配不同图像时特征对齐十分困难. 基于姿势提取方法的核心思想是充分利用行人的姿势信息, 增强对人体部位 (头、四肢、躯干等) 间的匹配.

在手工特征阶段, 就有方法开始考虑人体姿势信息. Farenzena 等[5] 提出了 SDALF 特征, 基于人体轮廓的对称轴区域将人体划分为头部、肢体、下肢三个部分, 靠近对称轴的像素将获得更高的权重分配, 从而抑制背景的干扰; Cheng 等[29] 使用了一种基于图形结构的细粒度的姿态表示来关注人体不同身体部分之间的对齐关系, 然后对每个部分的颜色特征进行更精确的匹配; Cho 等[30] 定义了前、后、左、右四个方位, 模型首先在不同方位下学习相应的人体匹配权值, 最后通过加权平均算法融合为多姿态匹配模型.

手工特征方法利用人体姿势信息划分重要区域, 能够获得表征能力更强的特征, 但因为缺乏更精确的姿势估计方法而进入研究瓶颈. 近年深度学习框架在姿势估计领域的成功[31−33], 以及不同研究方向之间的融合, 使得行人重识别领域也建立了一套基于深度学习姿势提取方法的基本架构, 包含三个步骤:

1) 计算人体的关键点: 使用姿势估计模型提取人体的关键点.

2) 获取行人特征: 设计特征提取网络提取行人特征图.

3) 特征融合: 将关键点信息与行人特征图结合, 实现人体部件的划分, 关键区域的对齐等.

Zhao 等[34] 首次将姿势提取模型应用与行人重识别任务, 网络 SpindleNet 的结构如图 5 所示. SpindleNet 包含区域生成网络 (Region proposal network, RPN)、特征提取网络 (Feature extraction network, FEN) 和特征融合网络 (Feature fusion network, FFN) 三个子结构, 分别完成人体关键点计算、行人特征提取和特征融合. 在 RPN 模块中, 首先由姿势估计模型得到人体的 14 个关键点, 然后生成表征人体部件的 7 个子区域: 头肩区域, 上身区域, 下身区域, 两个手臂区域和两个腿部区域. 特征提取网络 FEN 利用卷积提取原始图像的全局特征, 通过池化将全局特征与 RPN 输出的人体部件特征初步融合, 生成 7 个不同区域的局部特征. 最终在 FFN 中按照图 5 的方式重新组合局部特征, 输出融合 7 个人体局部区域的多尺度特征. SpindleNet 结合人体姿势信息进行预测, 能够有效避免姿态错位导致的特征对齐困难, 适合与在 大姿态变化的场景下使用, 在 Market-1501 数据集上取得了 rank-1 76.9%. 姿势估计模型因存在预测错误, 导致关键点提取不精确, Zheng 等[35]通过两层全连接层计算出人体部件的判别向量, 在对关键点特征融合时一起输入以校准姿势估计, 为姿势估计模型

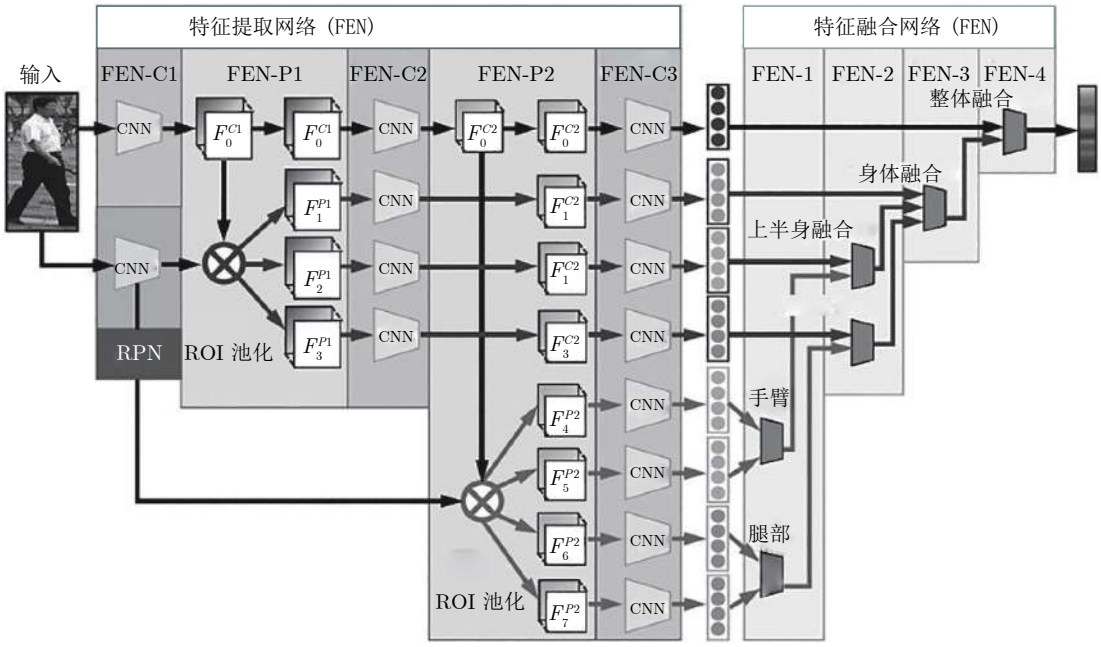


图 5　多分支融合姿态信息的 SpindleNet 网络流程图

Fig. 5　The pipeline of SpindleNet which fusions pose information with multiple branches

生成的人体特征提供了可靠的判别指标, 可挖掘更稳定的姿态对齐特征并减少估计模型的检测误差.

当行人重识别发生部分遮挡, 从整个图像中提取的特征会包含较多的噪声. 如果不对模型区分遮挡区域和人体区域, 易导致错误的检索结果. 为了抑制遮挡带来的噪声, 传统方法中一般通过手工剪裁去除图像的被遮挡部分, 然后输入未被遮挡的部分进行查询. 但面对数据规模较大的行人数据库时, 人工剪裁并不现实. 考虑到姿势提取模型对遮挡物体像素十分敏感, Miao 等[36] 使用姿势提取模型计算人体关键点的置信度. 在于行人特征融合时, 置信度较低的关键点相干的人体区域会被判定为遮挡区域, 其权重被置零从而不参加后续计算, 避免了手工处理遮挡区域.

3.1.2 基于生成对抗网络的姿势扩充

行人重识别目前缺乏大姿态变化的数据引导模型学习具有鉴别性的视点不变特征, 难以有效消除姿态变化对人体外观的影响. 而在近几年, 学术界常将生成对抗网络 (GAN) 应用于生成图像实现数据集的扩充. GAN 是由 Goodfellow 等[37] 在 2014 年提出的一种生成式模型, 基于 GAN 的数据扩充是图像风格迁移研究的应用, 其工作原理可概括为由生成模型学习目标图像的分布将特定的像素风格迁移到生成图片上, 鉴别模型判定生成图像的真伪, GAN 的目标就是不断训练直到鉴别模型相信生成图像, 达到以假乱真的效果. 文献 [16] 最早将 GAN 用于行人重识别的数据集扩充任务, 通过指定一个统一数据集的图像分布引导模型生成大量的无标签行人样本. 但是这种传统的无监督 GAN 存在以下问题:

1) 无监督学习不能充分利用样本的信息, 难以指导模型获得更好的判别能力

2) 传统的 GAN 方法只重视生成图像的视觉效果, 但由于人体形态复杂, 容易获取到严重扭曲的行人样本

针对这些问题, Liu 等[38] 引入姿势信息辅助 GAN 学习, 他们利用 MARS 数据集上提取的大量人体骨架特征构建一个引导模型, 同时将姿势信息输入到生成模型中校准生成图像, 最后通过整合人体关键点与目标数据集的外观得到带标签的姿势变换样本. 利用姿势信息辅助 GAN 的方法相较于传统方法, 生成图像更具真实感、位姿更可控. 以往的 GAN 模型只考虑生成的样本的真实性, 人体结构容易受破坏, Zhu 等[39] 采用多层级联的网络结构训练两个判别模型, 利用反向传播算法同时优化图像质量和骨架形态, 使生成的行人图像具有良好的锐利度一致性和外观一致性.

3.1.3 小结

表 2 归纳总结了本节的相关工作. 深度学习方法使用的基础网络一般为 GoogleNet[40]、ResNet50[41], 而在 GAN 方法中, 都使用了在像素对齐上表现优异的 CGAN (Conditional GAN)[42]. 随着姿态估计领域的发展, 模型的选用存在较大差异, 例如早期的 CPM (Convolutional pose machines)[43] 模型只能应用于单人姿态估计, 而 HPE (Human pose estimation)[44] 及 AlphaPose[45]可用于多人场景但计算耗时严重, 开源项目 OpenPose[46] 改进了 CPM 及 HPE, 可以达到实时的多人二维姿态估计. 总的来说, 结合姿势估计模型的行人重识别方法充分利用了骨架关键点, 将特征图转化为具有明确语义信息

表 2 基于姿势估计的方法总结 (rank-1 为原论文在 Market-1501 上的实验结果)

Table 2 Summary of pose estimation based methods (rank-1 refers to the result of original paper on Market-1501)

文献	来源	方法名称	基础网络或主要方法	方法类型	姿态估计		rank-1 (%)	主要工作概述
					模型	关键点数目		
[5]	CVPR10	SDALF	颜色相关图, 颜色矩	手工特征	—	—	—	设计颜色直方图等手工特征提取人体对称轴附近的局部信息.
[34]	CVPR17	SpindleNet	GoogleNet	深度学习	CPM	14	91.5	人体关键点定位人体部件ROI, 与行人特征级联融合生成鉴别性更强的特征.
[35]	Arxiv17	PIE	ResNet50	深度学习	CPM	14	78.6	双层全连接层提取人体部件判别向量, 指导姿态估计模型精确提取关键点.
[36]	ICCV19	PGFA	Resnet50	深度学习	AlphaPose	18	91.2	利用姿态估计模型对遮挡的敏感性预测遮挡区域, 降低遮挡对模型判别的影响.
[38]	CVPR18	Pose-transfer	CGAN	GAN	HPE	18	87.6	引入姿态估计模型定位人体结构, 优化GAN模型对人体形态的构建.
[39]	CVPR19	PATN	CGAN	GAN	OpenPose	18	—	采用双判别器分别改善图像质量及姿态形体, 提升生成图像的真实感.

的人体部件, 可以精确定位人体关键部位, 在解决行人姿态变化、物体遮挡问题上有显著的效果. 与传统的手工特征相别, 现有的姿势提取模型能够更精确提取关键点特征, 并且不局限于提取颜色特征, 通过深层次的特征融合保留了更多有效信息. 同时姿态信息与 GAN 结合可生成大量的多姿态行人数据, 在解决行人重识别缺乏跨视点配对的训练数据、大位姿变化下学习鉴别性特征和视点不变特征等重要问题上有着良好的发展前景. 现存的姿势估计方法的主要问题在于增加了模型的训练成本及网络结构复杂程度, 由于计算性能的限制, 在实际应用时需要降低姿势提取网络的额外开销. 在未来的研究中, 设计高效低耗的姿势提取和特征融合算法是结合姿势提取方法的研究重点.

3.2　基于特征空间分割

3.2.1　特征空间分割模型

与基于姿势提取的模型不同, 基于特征空间分割的方法不关注具体的人体语义部件, 而是在空间尺度上将特征划分为多个局部显著性区域, 让模型在训练过程中学习不同区域的差异性. 文献 [17] 中提出了一种利用图像水平分割提取人体抽象部件的 PCB (Part-based convolutional baseline) 模型, 网络结构如图 6 所示. 其主要思想是在共享卷积网络之后, 将获取到的全局特征在水平方向划分为均匀的 6 个区域, 代表人体的抽象部件特征. 由于多次卷积计算后特征具有良好的高层语义特征, 模型只需学习抽象部件之间的上下文关系. 这个简单高效的方法刷新了主流数据集上实验结果, 后续基于特征空间分割的研究大多以此为基准进行扩展. Wang 等[47] 认为 PCB 模型只关注了人体的细粒度特征, 忽视了整体对局部学习的影响, 提出一种整合全局特征和局部特征的多粒度模型 MGN (Multiple granularity network), 同时在特征约束上作出

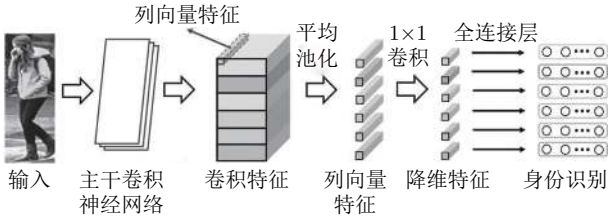


图 6　水平分割特征图的 PCB 网络

Fig. 6　The PCB network which partitions feature map horizontally

了改进, 将基于正负样本度量的三元损失[48] 用于优化全局特征, 分类能力强的交叉熵损失约束局部特征. 实验表明, 全局特征增强了模型对背景信息的学习, 而局部特征可融合人体关键特征, 与只使用局部特征的 PCB 相比, 模型识别精度有较大提升, 在 Market-1501 中提高了 5% 的 mAP. Zheng 等[49] 提出了由粗粒度到细粒度的渐近式金字塔模型, 不仅同时保留了全局特征和局部特征, 而且还考虑了它们之间的渐变关系. 最后融合的特征之间具有较强的上下文联系, 在复杂度较高的 MSMT17 数据集上取得了最领先的识别效果.

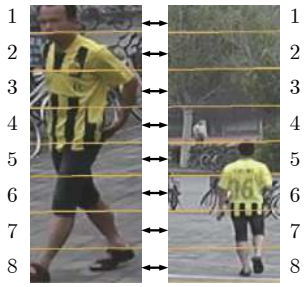
基于空间分割的方法能够简单直观的细化特征, 加强特征的鲁棒性, 却存在特征之间的对齐问题. 如图 7 所示, 由于距离与视角的不同, 人在图片中的相对位置不是固定的. 在这种情况下分割的局部特征之间无法被正确的匹配, 真实复杂场景会带来更大的误差.

为了解决这个问题, Luo 等[50] 提出了一种利用动态规划进行部件之间匹配的算法. 定义给定的一对行人特征 $F=\{f_1,\cdots,f_H\}$ 和 $G=\{g_1,\cdots,g_H\}$ 由多个局部特征的集合组成, 首先按照式 (5) 计算所有局部特征间距离 d_{ij} 的归一化表示.

$$d_{i,j}=\frac{e^{\|f_i-g_j\|_2}-1}{e^{\|f_i-g_j\|_2}+1},\quad i,j\in 1,2,3,\cdots,H \tag{5}$$

(a) 不同视角下的行人差异

(a) Pedestrian differences from different viewpoints



(b) 水平块的对齐错误

(b) The alignment error of horizontal stripes

图 7　视差导致的特征对齐问题

Fig. 7　The feature misalignment problem caused by parallax

最终的行人特征间相似距离由所有局部特征间的最短距离总和 S_{ij} 表示, 由式 (6) 中动态规划算法得到.

$$S_{i,j} = \begin{cases} d_{i,j}, & i = 1, j = 1 \\ S_{i-1,j} + d_{i,j}, & i \neq 1, j = 1 \\ S_{i,j-1} + d_{i,j}, & i = 1, j \neq 1 \\ \min\left(S_{i-1,j}, S_{i,j-1}\right) + d_{i,j}, & i \neq 1, j \neq 1 \end{cases} \tag{6}$$

该算法的特点在于考虑了特征匹配问题都存在一个自上而下的对齐约束, 只通过两组局部特征间的最短距离实现对齐, 而不需要人体姿势、视角等额外的监督信息.

通过解决特征对齐偏差, 空间分割方法在遮挡部位对其和跨域背景学习上有十分优异的性能. Sun 等[51] 在 PCB 的基础上继续探讨了如何在不增加先验语义信息的条件下赋予抽象部件更强的可解释性, 他们在输入图像中预先定义一个固定的部件分割, 将图像分割为上、中、下三个区域, 网络在自监督学习的过程中自动对每一个部件的可见性与不可见性打分, 分值低的区域被设置为不可见部件, 减少参与后续计算的权重. 由于固定部件的设置是可自行设置与可自动获取, 网络能够指导性地区分有效和无效的部件特征, 该方法能够有效地解决遮挡问题, 在 Partial-REID 遮挡数据集上取得了最优结果. 为了解决行人重识别的跨域偏差问题, Fu 等[52] 利用局部分割特征挖掘目标域样本中的自然相似特征, 他们将无标签的目标域数据集中的所有人划分为全身、上半身、下半身三组, 无监督地将目标域的同一个行人按照 3 个分组聚类, 每个人都可以根据所属的分组分配一个伪标签, 由此构建一个新的子数据集细调网络. 与以往基于数据扩充等处理跨域识别的方法相比, 利用局部特征能够使模型学习背景和人体部件的差异性, 并且进行无监督学习不需要额外标注目标数据集, 有很好的实际应用价值.

3.2.2 小结

表 3 归纳总结了本节相关工作. 从表中可以看出, 以 ResNet50 为基准网络, 使用交叉熵损失和三元损失同时约束网络训练是目前特征分割方法的常用结构. 在网络复杂度方面, 特征分支的增加会加大网络的计算量及收敛难度, 但从实验结果来看, 全局特征与局部特征结合、合理地划分更多特征都能够为网络性能带来提升. 总的看来, 基于图像水平分割的行人重识别方法, 能够简洁有效地获取局部特征, 并通过不同的特征组合方式, 使模型的性能显著性地提高, 屡次刷新公开数据集的最好识别效果. 目前基于特征空间分割的算法仍存在三个需解决的问题:

1) 可解释性差. 对图像的分割都是在多层卷积网络之后进行的, 分割的正确性难以被验证.

2) 训练效率低. 通过特征分割会得到多个局部特征, 目前的训练策略是对每一个特征都单独优化, 随着特征划分数量的提升, 训练难度和资源消耗也极大的增加.

3) 特征对齐难. 虽然目前已经有很多针对特征对齐问题的研究, 探索更具有指导性的特征对齐方法仍是研究的重点.

在未来的研究中, 设计更合理的分割与对齐方式、更有效的特征融合方法、和更高效训练策略是特征空间分割方法研究的发展方向.

表 3 基于特征空间分割的方法总结 (rank-1 为原论文在 Market-1501 上的实验结果)

Table 3 Summary of feature spatial partition based methods (rank-1 refers to the result of original paper on Market-1501)

文献	来源	方法名称	基础网络	损失函数	分割数目统计		rank-1 (%)	主要工作概述
					全局特征	局部特征		
[17]	ICCV18	PCB	ResNet50	交叉熵损失	0	6	93.8	提出水平分割卷积特征, 提取细粒度的局部特征.
[47, 53]	ACM19	MGN	ResNet50	交叉熵损失 三元损失	3	5	95.7	多粒度网络, 结合粗粒度的全局特征及细粒度的局部特征, 使用多损失联合训练.
[49]	CVPR19	Pyramidal	ResNet50	交叉熵损失 三元损失	1	20	95.7	构建金字塔结构, 在分割特征的同时保留特征间的上下文关系.
[50]	PR19	AlignedReID	ResNet50	交叉熵损失 三元损失	1	7	91.8	设计了一种动态规划算法, 优先匹配相似度更高的局部特征, 减少了特征对齐误差.
[51]	CVPR19	VPM	ResNet50	交叉熵损失 三元损失	0	3	93.0	预定义分割区域, 使特征分割模型更稳定的提取部件特征.
[52]	ICCV19	SSG	ResNet50	交叉熵损失 三元损失	0	3	86.2	与无监督学习结合, 将每个分割区域作为一类聚类中心, 构建目标域与原域的细粒度相关性.

3.3　基于视角信息

现有数据集的图像多是水平拍摄, 缺乏显著的视角变化. 但在真实场景下行人相对拍摄点的姿势及方向会极大地影响人体视觉外观. 例如俯视、仰视易使行人特征不完整, 侧视图与直视图下的行人外观也存在巨大差异. 目前处理视角姿态变化的方法主要依赖于姿态关键点或相机方向信息, 无法充分利用行人在不同视角下的特性. 本小节将重点论述视角信息在行人重识别任务中的影响与应用.

3.3.1　视角对行人重识别的影响的研究

真实场景中, 视角的不同会导致人体姿态不一致, 增大了行人图像的复杂程度. 但若对聚合不同观察角度特征的视角信息加以利用, 可显著提升模型在复杂场景下的泛化性能. 为了研究视角的影响, Sun 等[54] 开发了一个大型的合成数据引擎 PersonX. PersonX 基于 Unity 开发, 构建了一个环境可控制的 3D 世界, 包括 1 266 个人的模型. PersonX 中的人物和物体贴近真实场景, 视角变量中的光源, 场景和背景都可编辑, 具有高度的灵活性与可调整性. PersonX 模拟了真实世界中视角对行人重识别精度的影响, 并构建出多姿态的行人测试数据, 以此为实验对象, 在 PCB 网络上进行实验, 得出如下结论:

1) 在数据集中缺失视角信息会影响模型的性能.

2) 缺失角度连续的视角比缺失随机视角影响更大.

3) 行人侧视图 (左、右视角) 的识别精度高于正视图 (前、后视角).

该研究构建了一个虚拟 3D 模型, 为现实世界数据集中光照、视角等环境变量难以设置的问题提供了一个有效的解决方案, 系统地研究了视角对行人重识别的影响, 为后续的视角研究提供了可靠的实验依据. 但是由于现实世界仍比虚拟模型情况更复杂, 测试数据集需赋予更多的视角变化, 使实验内容更贴近真实世界.

3.3.2　通过视角方法改进行人重识别

现有的大多数算法都建立在视角不变的假设上, 忽略了身体对齐问题. 图 8 给出了行人外观在不同摄像头下的差异, 图 8(c) 的特征可视化结果也表明现有模型在不同视角下特征关注点也不同. 为了对不同视角下的人体特征变化加以利用, 传统手工特征多通过多视角建模的提取角度特征. Bak 等[55] 认为人体位置的微小变化可根据人体的对称性和不对称性确定, 距离物体对称轴较近的特征能够更好地抵抗背景噪声. 首先通过仿射变换将平面图像旋转到与地面正交, 然后利用不同的位姿方向信息一个构建三维场景, 划分不同对称簇改善不同视角下的图像对齐效果. 该方法在不使用深度学习的情况下取得了较好的效果, 在早期的小型数据集[56]上取得了最优识别率. 考虑到不同视角图像间具有较强的相关性, Wu 等[57] 提出了一种新的姿态先验技术, 首先通过大量前视图和后视图数据估算出对应视角的角度描述符, 然后以描述符作为先验依据计算待检测行人的方位置信度, 以置信度水平评价人体姿态相对于摄像视角的变化程度, 进而估算人体的准确姿势, 有效地解决了大位姿变化场景下的行人对齐问题.

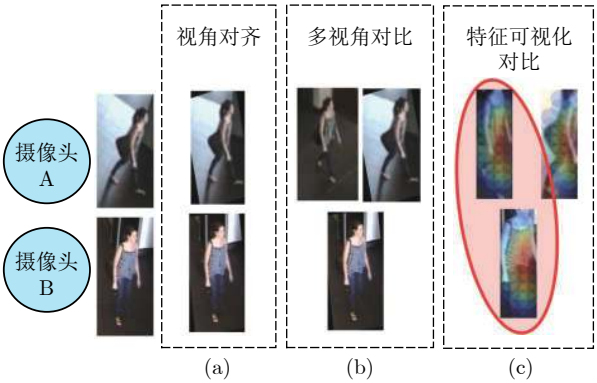


图 8　不同摄像头采集的行人特征对比示例

Fig. 8　The comparison of pedestrian feature representation captured by different cameras

早期的手工方法提取视角特征在大数据量的情况下缺乏良好的泛化能力, Li 等[58] 在 2016 年提出了行人属性数据集 RAP, 对行人的视角信息进行了标注, 为基于深度学习的视角特征研究方法提供了实验数据. Sarfraz 等[59] 利用 RAP 数据集中的视角标注训练出视角估计模型 VeSPA (View-sensitive pedestrian attribute) , 可预测行人前、侧、后视角的概率. 在 CVPR2018 中, Sarfraz 等[60]构建了 PSE (Pose sensitive embedding) 模型, 可通过 3 个视角单元分别学习训练数据中前向、侧向、后向视角的特征, 如图 9所示. 3 个视角单元会共享前端网络所提取的全局特征, 然后使用预训练的 VeSPA 模型在视角分支中计算行人视角概率, 视角概率值会作为对应单元的权重得到最终的加权融合特征. 该方法端到端地结合了行人视角信息, 在视角多变的场景下具有较好的效果.

3.3.3　小结

表 4 中归纳总结了本节的相关工作, 其中手工特征的目标是通过角度描述符、位姿坐标校准等方法 提取视角不变特征, 而深度学习方法则利用大量数据构建视角估计模型, 在近年来也涌现出了更新

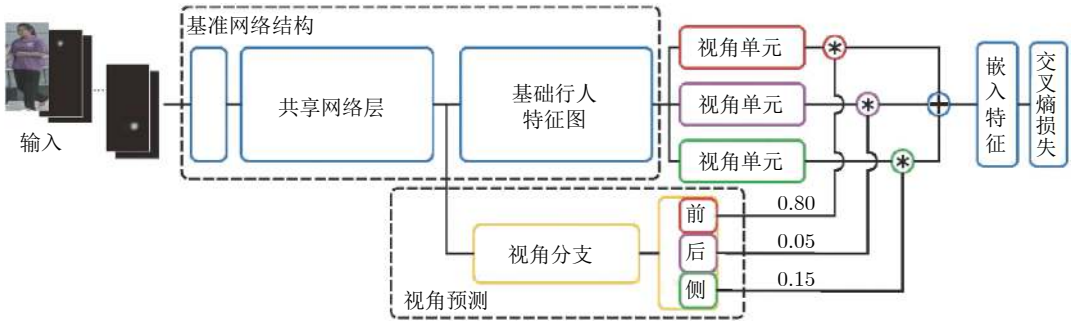


图 9 结合视角估计模型的 PSE 网络

Fig. 9 The PSE network which combines viewpoint estimation model

表 4 基于视角信息的方法总结

Table 4 Summary of viewpoint based methods

文献	来源	基础网络或主要方法	方法名称	损失函数	方法类型	主要工作概述
[54]	CVPR19	PCB	PsrsonX	交叉熵损失	深度学习	提出了一个3D行人数据集, 定量探讨了视角特征对行人重识别任务的影响.
[55]	AVSS14	坐标仿射变换	TA + MS + W特征	—	手工特征	挖掘人体对称性特征、角度特征, 利用仿射变换对齐图像.
[57]	TPAMI14	角度描述符	VIH	—	手工特征	多视图构建角度描述符, 预测固定摄像头下行人姿态变化情况.
[59]	BMVC17	GoogleNet	VeSPA	交叉熵损失	深度学习	基于行人属性集的视角标注, 训练了一个分类模型, 可预测行人视角概率.
[60]	CVPR18	ResNet50	PSE	交叉熵损失	深度学习	将VeSPA模型用于行人重识别任务, 结合视角概率值生成鉴别特征.

颖的研究思路. 总的来看, 基于视角信息的方法将影响行人重识别效果的视角因素转变为有益模型学习的先验特征, 用以校正模型对原始特征的判断, 在光照变化复杂、行人位姿变化大的场景下都有较稳定的识别精度. 但由于目前数据集含有的多视角样本数量稀少, 多视角建模比较困难, 对视角信息的研究欠缺 定量、稳定的分析手段. 目前在 PersonX 中给出了以 3D 建模模拟真实世界的思路, 在未来的研究中, 可以在此基础上优化模拟视角的有效性.

3.4 基于注意力机制

在图像领域, 注意力机制的原理类似于人类观察物体时将视线聚焦于小部分重要区域. 对于行人重识别而言, 注意力机制的目标就是寻找对特征图影响较大的区域, 增强模型对重要局部区域的关注度. 注意力机制不需要人工手动标注关键区域, 可实现端到端的行人重识别任务, 同时在大数据驱动之下也具有良好的泛化能力. 本小节将注意力机制归纳为为空间注意力、通道注意力及部分较特殊的非卷积注意力方法, 并论述各类代表方法.

3.4.1 空间注意力

空间注意力的工作原理可以概括为图 10, 首先通过局部建模从原始特征的空间像素分布中学习一个的注意力图 (Attention map), 注意力图将特征中不同平面区域对关键信息对贡献大小数值化, 然后在训练过程中与原始特征融合得到局部加权的特征图, 通过反向传播算法自主更新重要区域的权重.

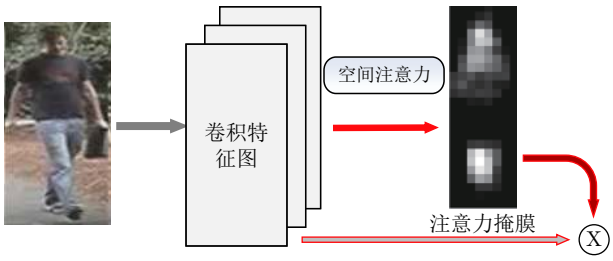


图 10 空间注意力机制方法工作原理示意图

Fig. 10 Illustration of spatial attention mechanism

考虑到不同卷积核对人体部件区域的关注度不同, Zhao 等[61] 提出由一个卷积层构成的人体部件区域检测器, 将全局特征图输入到多路检测器分支中, 最后在每个分支都能产生不同区域的注意力加权图. 与当时的特征空间算法相比, 基于空间注意力的方法不易受特征对其问题的影响, 在 Martket-1501 和 CUHK03 数据集上都取得了更优秀的结果. 空

间注意力机制可以突出卷积网络的有效部分并抑制背景噪声的干扰, Song 等[62] 结合图像分割, 分离图像的前景与背景, 将人体轮廓设置为二值掩膜, 然后设计了由轮廓掩码引导的注意力模型来分别学习身体和背景区域的特征. 并使用区域级的三元损失学习全局特征与身体局部特征间的关系, 及与背景像素的差异, 有效抑制了像素级的背景杂波.

3.4.2　通道注意力

通常输入到卷积网络中的图像初始由 R、G、B 三通道表示, 在经过不同的卷积核之后会产生更多的特征通道, 每一个通道都表示参与计算卷积核的不同分量, 因此对整体图像的权值贡献各不相同. 通道注意力机制的目的就是将每个通道的权重压缩到一个特征向量中, 最后与原始特征融合 使模型学习关键通道的特征, 其原理类似与信号的时域变化, 如图 11 所示.

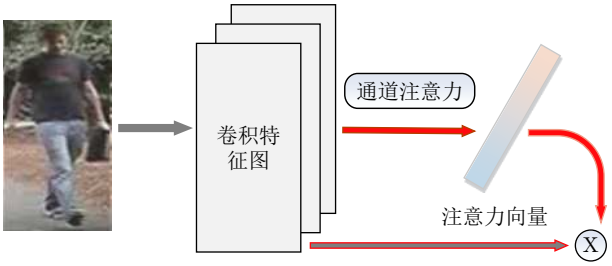


图 11　通道注意力机制工作原理示意图

Fig. 11　Illustration of channel attention mechanism

在行人重识别任务中的通道注意力模块普遍使用通道压缩−权重激活−重整权重 (Squeeze-excitation-reweight, SER) 结构, 最早由 Hu 等[63] 提出. 首先在式 (7) 中用全局平均池化每个通道中的特征值相加后平均, 使每个通道的特征都被压缩到一个 1×1 的二维矩阵中, 由二维矩阵表征特征通道上所响应的全局分布.

$$z_c = F_{sq}(u_c) = \frac{1}{H \times W}\sum_{i=1}^{H}\sum_{j=1}^{W} u_c(i,j) \tag{7}$$

式中, z_c, u_c 分别表示输入 u、输出特征 z 的第 c 个通道.

然后使用激活函数 ReLU[64] 和 Sigmoid 按照式 (8) 激励通道特征图, 即使特征图先经过两层全连接层, 而后利用激活函数加强通道特征的非线性响应, 得到最终的通道注意力图.

$$s = \sigma(g(z,W)) = \sigma\left(W_2\delta\left(W_1 z\right)\right) \tag{8}$$

其中, δ 表示激活函数 ReLU, σ 表示激活函数 Sigmoid, W_1 和 W_2 为全连接层的权重矩阵.

最后利用式 (9) 融合原始特征与通道注意力图, 使原特征图中的各个通道的权重被重新分配.

$$\widetilde{x}_c = F_{\text{scale}}(u_c, s_c) = s_c u_c \tag{9}$$

在 CVPR2018 中提出的 HA-CNN (Harmonious attention convolutional neural network)[65] 结合了空间注意力及通道注意力, 并通过多尺度加权融合特征, 将粗糙的区域注意力与通道细节信息结合, 提高了注意力选择与特征表示的兼容性. 考虑到一些特征通道共享前景人物、遮挡或背景等相似的语义上下文, Chen 等[66] 认为特征通道也存在类别差异, 使用 softmax 算子将通道注意力图转换为一个二维矩阵, 增强了相同类别通道间的关联性, 最后联合全局特征检索行人, 在 Market-1501 数据集上达到 88.28% mAP 的目前最优结果.

3.4.3　非卷积注意力方法

注意力机制的本质是加强模型对关键区域的关注程度, 有部分方法不使用简单的卷积计算学习特征的权值分布, 而是从注意力机制的本质出发展开研究. Fu 等[67] 提出 “Look closer to see better”, 在训练过程中对特征局部区域进行尺度放缩, 选择放大其中的关键区域, 并将该区域以半监督学习的方法再次输入网络参与训练. 该方法利用注意力机制在训练过程中扩充训练样本, 在行人重识别等细粒度分类任务中都有较好的表现. Dai 等[68] 认为局部特征学习的本质是降低网络对显而易见的全局特征的关注度, 设计了一种批次特征擦除网络 BDB, 对于同一输入批次的特征图随机遮挡一块区域, 强迫网络学习未遮挡区域的细节特征. 该模型只通过嵌入简单的数据增强技术, 便在 Market-1501 数据集上取得了 95% 的 rank-1.

3.4.4　小结

表 5 归纳总结了基于注意力机制的方法, 可以看出该类方法的精度在近几年取得较大的提升, GoogleNet、ResNet50 等基础网络一般在 ImageNet 数据集上进行预训练, 而 MGCAM (Mask-guided contrastive attention model) 及 HA-CNN选择自搭的 CNN 网络作为基础网络, 其优点是可以针对输入数据从头训练网络. 总的来说, 注意力机制方法通过压缩人体关键特征, 使模型自主学习图像中的重要信息, 增强了模型对噪声的抗干扰能力. 其中基于空间注意力机制的方法, 主要是通过卷积将原始图像的空间信息压缩, 过滤掉对模型判别性能影响较小的无关信息; 而基于通道注意力的方法融合了卷积特征的多通道信息, 不仅在行人重识别任务上有较好的表现, 同样与计算机视觉的其他领域有较好的适应性, 被广泛地整合到 GoogleNet、ResNet 等重要的基础网络模型中. 部分非卷积的注意

表 5 基于注意力机制的方法总结 (rank-1 为原论文在 Market-1501 上的实验结果)
Table 5 Summary of attention based methods (rank-1 refers to the result of original paper on Market-1501)

文献	来源	方法名称	基础网络	实现方法	损失函数	方法类型	rank-1 (%)	主要工作概述
[61]	CVPR17	DLPAR	GoogleNet	多分支的 1×1 卷积层	三元损失	空间注意力	64.2	利用多个注意力模块作用到不同的人体部件, 多分支提取鉴别性特征.
[62]	CVPR18	MGCAM	MGCAN	全卷积网络[69]	交叉熵损失 三元损失	空间注意力	83.7	结合背景分割, 提取二值化轮廓图作为注意力图, 降低杂乱背景的干扰.
[65]	CVPR18	HA-CNN	CNN	SER结构结合多层卷积	交叉熵损失	空间注意力 通道注意力	91.2	融合空间注意力学习与通道注意力, 同时学习平面像素特征与通道尺度特征.
[66]	ICCV19	ABD-Net	ResNet50	Softmax层加权特征矩阵	交叉熵损失 三元损失	空间注意力 通道注意力	95.6	利用 softmax 的分类特性, 加强通道特征间的相关性.
[68]	ICCV19	BDBNet	ResNet50	DropBlock[70] 层改进	交叉熵损失 三元损失	非卷积方法	95.3	特征正则化, 将随机擦除作用到特征学习, 可有效抑制过拟合.

力方法将数据增强操作作用在特征的学习过程中, 对特征进行擦除、剪裁等操作, 同样能够达到较领先的效果. 目前基于注意力机制的方法仍存在以下难点:

1) 注意力选择的可解释性: 由于卷积网络欠缺良好的可解释性, 目前只能通过一些反卷积[71] 或是反向传播[72] 的方法观察特征图的权值分布, 导致难以指导卷积网络精确提取目标特征, 且无法确定注意力区域关注更多的是颜色特征还是轮廓特征. 在未来的研究中, 应该探索可验证、指导性更强的 注意力算法.

2) 注意力机制的作用域: 现有方法都以特征整体为参照对象学习权重图, 使注意力机制的作用域为图像整体, 通常只能从全局挖掘最显著的区域特征, 导致模型容易忽略部分关键的局部细节, 因此如何约束注意力机制的作用域以学习更充分的局部信息也是一个有意义的研究方向. 结合目前的研究成果, 可以探索注意力机制与其他局部特征结合所形成的局部注意力方法, 而随着注意力作用局部区域的增加, 如何减少计算量及维持不同区域之间权重平衡也是值得关注的研究内容.

4 局部特征的行人重识别方法实验对比与分析

4.1 典型算法对比

前面主要介绍了行人重识别领域中常用的公共数据库和近些年来提出的相关识别方法及其发展. 本节将结合 DukeMTMC-reID、Market-1501、CUHK03 三个主流数据集具体对比分析相关识别模型.

本节在近几年顶级会议及顶级期刊上发表的行人重识别方法中, 首先挑选了基于局部特征的各类方法中准确率较高且具有代表性的方法. 为了便于比较, 我们还选择了 具有代表性的传统手工方法、基于全局特征的方法及无监督学习方法, 在表 6~8 按照传统手工方法、基于深度学习的全局特征方法与局部特征方法分为了 6 组, 并统计了代表方法在三个数据集中的表现, 表中的所有实验结果均由原论文给出.

从表 6~8 的对比数据中可以看出, 在三个主流数据集中, Market-1501 数据集上的拟合程度最高, 一选准确率最高超过了 95%, 而 CUHK03 是目前最具有挑战性的数据集, 所有方法在该数据集上的实验结果都相对较低, 而基于深度学习模型的效果相较于传统手工方法得到了大幅度的提升, 证明深度学习方法在大规模数据下有更好的泛化能力. 在无监督学习方法中, 在 CVPR2018 以前行人重识别领域正式发表的无监督学习方法只有 UMDL (Unsupervised cross-dataset transfer learning)[73], 基于不变性字典学习, 而以 SPGAN (Similarity pre-

表 6 DukeMTMC-ReID 数据集上各种方法的对比结果 (%)
Table 6 Experimental results of various methods on DukeMTMC-ReID dataset (%)

方法	类型	rank-1	mAP
XQDA + LOMO[10] (2015)	手工特征	30.7	17.0
UMDL[73] (2016)	无监督 + 手工特征	30.0	16.4
SPGAN[74] (2018)	无监督 + GAN	46.9	26.4
PAN[1] (2017)	全局特征	71.5	51.5
Pose-transfer[38] (2018)	姿势提取	78.5	56.9
MGN[47] (2018)	特征空间分割	88.7	78.4
Pyramidal[49] (2019)	特征空间分割	**89.0**	**79.0**
PSE[60] (2018)	视角信息	79.8	62.0
HA-CNN[65] (2018)	注意力机制	80.5	63.8

表 7　Market-1501 数据集上各种方法的对比结果 (%)
Table 7　Experimental results of various methods on Market-1501 dataset (%)

方法	类型	rank-1	mAP
XQDA + LOMO[10] (2015)	手工特征	43.8	22.2
UMDL[73] (2016)	无监督 + 手工特征	34.5	12.4
SPGAN[74] (2018)	无监督 + GAN	58.1	26.9
SOMAne[3] (2017)	全局特征	73.9	47.9
Spindle[34] (2017)	姿势提取	76.9	—
Pose-transfer[38] (2018)	姿势提取	87.6	68.9
PCB[17] (2018)	特征空间分割	92.3	77.4
MGN[47] (2018)	特征空间分割	95.7	86.9
Pyramidal[49] (2019)	特征空间分割	**95.7**	**88.2**
PSE[60] (2018)	视角信息	87.7	69.0
HA-CNN[65] (2018)	注意力机制	91.2	75.7
ABD-Net[66] (2019)	注意力机制	95.6	**88.2**

表 8　CUHK03 数据集上各种方法的对比结果 (%)
Table 8　Experimental results of various methods on CUHK dataset (%)

方法	类型	rank-1	mAP
XQDA + LOMO[10] (2015)	手工特征	12.8	11.5
PAN[1] (2019)	全局特征	36.3	34.0
Pose-transfer[38] (2018)	姿势提取	41.6	38.7
PCB[17] (2018)	特征空间分割	61.3	54.2
MGN[47] (2018)	特征空间分割	**66.8**	**66.0**
HA-CNN[65] (2018)	注意力机制	41.7	38.6

serving generative adversarial network)[74] 为代表的深度学习无监督方法较基于手工特征的 UMDL 识别精度显著提高, 但由于不使用数据集的标注信息, 与其他有监督方法相比仍有较大的提升空间. 与基于全局特征的方法相比, 三个数据集的实验结果都表明基于局部特征的方法具有更优秀的识别性能.

表 9 中总结了四类局部特征方法在特征学习上的差异, 其中视角信息与姿态提取的方法都需在学习过程中融合准确的语义信息, 由于姿态估计等数据集与行人重识别数据集有较大的域偏差, 估计模型泛化能力不足, 识别精度仍有待提高. 而特征空间分割方法和注意力方法均没有使用具体的语义特征, 在三个主流数据集上都有较优的表现.

表 10 展示了将多种局部特征融合的现有方法在 DukeMTMC-reID 数据集上的实验结果. PGFA (Pose guided feature alignment)[35] 在 PCB 的基础上增加了一个姿势估计分支, 利用关键点特征消除遮挡区域, 最后与 PCB 分支所提取的特征联合预测. 论文在数据集检索图像被局部遮挡的情况下进行了遮挡实验, 试验结果表明 PGFA 在遮挡场景下抗干扰能力有较大的提升. P^2 -Net

表 9　各类局部特征方法比较
Table 9　Comparison of various local feature methods

方法类型	对应文献	特征学习特点	影响性能的主要因素
姿势估计	[5, 29–39]	在特征学习的过程中融合准确的关键点特征, 以学习更具鉴别性的特征, 或利用关键点处理人体定位对齐、遮挡问题.	姿态估计模型对人体关键点的检测精度、特征融合方法的有效性. 姿态估计数据集与行人重识别数据集具有较大偏差, 造成姿态估计模型在行人重识别任务中的语义分割效果不佳.
特征空间分割	[15, 47–52]	对卷积层的特征进行均匀分割, 生成的每一块特征都由单独的损失函数约束训练	输入数据的复杂程度, 特征分割区域的稳定性, 易受局部特征对齐问题的影响, 依赖质量较高的数据.
视角信息	[54–60]	需要准确的视角信息. 常利用视角信息对不同视角的图像进行仿射变换以对齐图像视角, 或融合视角信息增加特征的鉴别性.	视角信息的准确性, 目前没有专门增对视角特征的研究领域且相关数据集较少, 视角估计模型的准确度还有待提升.
注意力机制	[61–68]	学习由卷积计算生成的显著性区域, 在训练过程中提高相关程度较高区域的权重, 同时降低相关程度较低区域的权重.	注意力选择的有效性及多样性, 相关的工作表明结合多类注意力机制能够获得更好鉴别性特征.

表 10　DukeMTMC-reID 上融合多类局部特征方法的实验结果 (%)
Table 10　Experimental results of the multiple-local feature fusion methods on DukeMTMC-reID (%)

方法	文献出处	类型描述	rank-1		mAP	
			原始数据	遮挡处理	原始数据	遮挡处理
PCB[17]	ECCV 2018	特征空间分割	81.9	42.6	65.3	33.7
PGFA[36]	ICCV 2019	特征空间分割+姿势估计	82.6	51.4	65.5	37.3
P^2 -Net[75]	ICCV 2019	特征根据分割+注意力机制	86.5	—	73.1	—

(Dual part-aligned network)[75] 考虑到现有语义分割模型在行人重识别任务上精度较低, 使用一个注意力分支粗糙的估计行人的语义部件, 同样与 PCB 分支所提取的特征融合, 在精度上有较大的提升. 总的来说, 融合多种局部特征可加强特征的鉴别能力, 其中影响姿势估计或视角信息与其他特征融合的主要因素是估计模型在行人重识别数据上的语义分割精度, 而注意力机制本身在图像领域有较高的泛用性, 易嵌入到网络模型的子结构中, 主要需要解决的问题有:

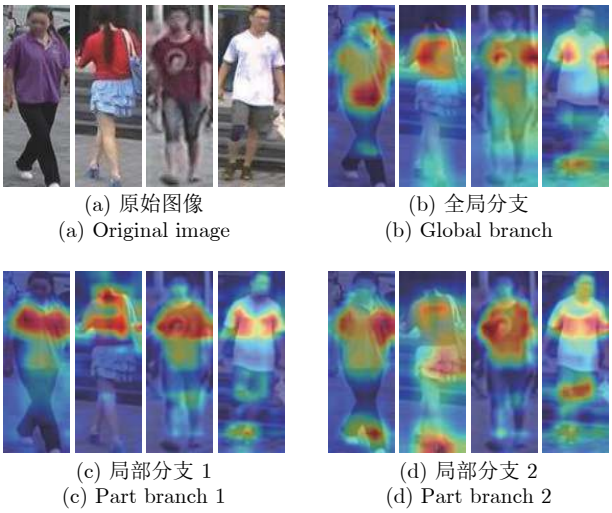
1) 增加注意力模块会使模型更加复杂, 造成训练难度及计算量增加.

2) 当多个区域都有独立的注意力模块作用时, 可能会破坏区域之间的相关性, 需要重新计算区域之间的权重比例.

初始空段落

4.2 全局特征与局部特征对比

为了进一步验证局部特征和全局特征关系. 图 12 给出了特征空间分割方法 MGN[46] 模型中两个局部特征分支和一个全局特征分支在 Grad-CAM (Gradient-weighted class activation mapping)[76] 算法下的特征图可视化结果. 图 12(a) 为原始输入图像, 图 12(b) 展示了全局特征分支的特征可视化结果, 特征的权重分布分散并且没有结构清晰的轮廓, 图 12(c) 和图 12(d) 为两个局部特征分支的特征可视化结果, 展示了网络关注的重点区域. 首先可以发现, 无论是全局特征还是局部特征, 都对人体上半身的关注程度较高, 而该区域含有服装所提供的色彩信息以及鲜明的人体轮廓信息, 说明该模型对颜色、轮廓特征关注程度较高. 相较于全局特征分支, 局部特征分支在重点区域的轮廓更完整清晰, 同时对关注区域的权重分配也更加集中, 明显优于全局特征分支提取的特征.



(a) 原始图像
(a) Original image

(b) 全局分支
(b) Global branch

(c) 局部分支 1
(c) Part branch 1

(d) 局部分支 2
(d) Part branch 2

图 12 MGN 中不同分支的特征可视化结果

Fig. 12 The feature visualization results of the different branch of MGN

5 目前工作的研究难点与展望

行人重识别研究已经持续了多年, 在现有的公开数据集上, 各类算法都取得了很大的突破. 本文所介绍的基于局部特征的行人重识别方法, 对性解决了人体姿态变化、视角差异、物体遮挡等研究问题. 但面对真实场景的海量视频数据, 面向局部特征的行人重识别技术要从研究到广泛的实际应用, 还面临一系列困难.

1) 局部特征对齐. 在现有的局部特征方法中, 只有基于姿势提取的方法能够有效处理特征间对齐. 其他局部特征缺乏指导性的特征对齐算法, 在姿态、视角变化更复杂的场景下对应部位的特征间易出现错位, 导致特征匹配受到严重影响.

2) 局部特征的可解释性. 目前的局部特征提取都基于深度学习, 随着网络层数的加深、模型结构的优化以及机器计算能力的提升, 虽然能够通过实验结果验证方法正确性, 但设计结构的理由、特征图包含的信息缺乏理论性的证明.

3) 训练困难. 通过特征空间分割以及姿态估计方法提取的所有局部特征都会被单独分配一个目标函数进行训练, 随着特征数量的增加, 其计算资源的消耗也将显著增大. 并且不同的目标函数通过反向传播共同优化模型, 数量越多越难维持目标函数间的平衡, 易造成模型训练无法收敛.

4) 特征提取与特征融合算法复杂度高. 现有的方法为了利用更多的局部信息都设计了较复杂的特征提取网络, 且较多局部特征的融合也会消耗大量的计算资源, 要做到实时的行人检索十分困难.

综上所述, 基于局部特征的行人重识别方法的分析与研究还处于逐步成熟阶段, 尚有广阔发展空间, 有很多困难需要我们去探索和认识. 对未来研究的可能发展趋势, 可以从以下几个方面进行考虑:

1) 结合自然语言的跨模态行人检索. 在部分行人重识别任务中, 存在无法获取到搜查对象的图像的情况, 例如在安防领域, 通常依靠目击者的语言描述搜索特点嫌疑人. 常规方法因为失去参考对象而无法使用, 而自然语言技术可将局部的行人属性特征[77] 提炼为描述性信息, 通过视觉信息与自然语言的多模态信息融合, 可以有效地解决信息模糊的问题.

2) 结合 GAN 生成针对性数据集. 本文所论述

的局部特征方法在一定程度上解决了行人重识别目前所面临的大位姿变化、物体遮挡等主要挑战, 同样可以从数据角度出发, 解决特定研究的数据稀缺问题. 如文献 [38] 中将人体姿态与 GAN 结合生成了大量多姿态行人数据, 有效避免了采集数据的困难. 在未来, 同样可以利用局部特征与 GAN 生成针对遮挡、多视角等问题的数据集, 可有效推动研究进展.

3) 基于无监督学习的特征匹配. 行人重识别的数据标注代价较高, 且目前的自监督学习模型在跨域应用上存在局限性, 每次更换目标检测域都需要重新调整. 而无监督学习有较好的泛化能力且不需要数据标签, 具有极高的性价比, 在学术界和工业界都是重要的研究方向. 同时无监督学习也为局部特征对齐研究提供了一个新的思路, 利用无监督学习为不同区域的局部特征生成伪标签, 使属于同一个类别的特征最终聚集在一起, 进而寻找最优匹配结果或筛选难对齐样本.

4) 构造更准确的唯一性特征. 对于现有的行人重识别方法而言, 行人衣装是极其重要的显著性特征, 在 TPAMI 的一篇工作[78] 中重点探讨了换衣重识别问题, 表明在行人服装发生显著变化的场景中, 现有方法的有效性会大幅下降. 而在夜间重识别及红外图像重识别等跨模态场景中, 通常只能获取大致的轮廓特征, 无法提取有效的行人衣装特征. 因此, 构造更具唯一性的特征也是未来的研究重点, 例如更准确的姿态及视角特征、行人步态特征、时序特征等语义级特征都值得更深入的研究.

5) 与细粒度图像识别领域结合. 行人重识别任务与细粒度图像识别的目标都是精确地划分相应数据中的子类别, 而细粒度图像识别发展较为成熟, 目前有大量具有启发性的工作值得行人重识别研究借鉴. 例如在 ICCV2019 的一篇工作[79] 将深度学习可解释性领域的工作与注意力机制结合, 利用反向传播在训练过程中提供了更具指导性的注意力选择区域. 而一些研究工作[80−81] 探讨了局部注意力 (Local attention) 的有效性, 认为特征不同通道的响应可作为近似的语义分割部件. 在 IJCV2020 的一篇文章[82] 中首次提出了细粒度行人重识别的概念, 定义为相似着装下的行人重识别任务, 论文中将姿势估计方法与循环神经网络结合, 生成更具鉴别性的动态姿态特征, 对后续工作有一定的启发性.

6 结语

行人重识别是计算机视觉领域的一个热门研究方向, 具有极大的实际应用价值. 针对基于局部特征的行人重识别方法, 本文首先介绍了现有的主流数据库, 然后分别从基于姿势提取、基于特征空间分割、基于视角信息 和基于注意力机制的四类方法对现有的研究成果进行了综述, 以实验数据为佐证阐明了各种方法的优缺点. 最后, 针对现有研究中的缺陷进行论述, 指出了该领域仍待解决的问题, 并深入探讨了未来发展的方向.

References

1 Zheng Z D, Zheng L, Yang Y. Pedestrian alignment network for large-scale person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*, 2019, **29**(10): 3037−3045

2 Chen H R, Wang Y W, Shi Y M, Yan K, Geng M Y, Tian Y H, et al. Deep transfer learning for person re-identification. In: Proceedings of the 4th International Conference on Multimedia Big Data (BigMM). Xi′an, China: IEEE, 2018. 1−5

3 Barbosa I B, Cristani M, Caputo B, Rognhaugen A, Theoharis T. Looking beyond appearances: Synthetic training data for deep CNNs in re-identification. *Computer Vision and Image Understanding*, 2018, **167**: 50−62

4 Gray D, Tao H. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In: Proceedings of the 10th European Conference on Computer Vision. Marseille, France: Springer, 2008. 262−275

5 Farenzena M, Bazzani L, Perina A, Murino V, Cristani M. Person re-identification by symmetry-driven accumulation of local features. In: Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Francisco, USA: IEEE, 2010. 2360−2367

6 Bazzani L, Cristani M, Murino V. Symmetry-driven accumulation of local features for human characterization and re-identification. *Computer Vision and Image Understanding*, 2013, **117**(2): 130−144

7 Lowe D G. Object recognition from local scale-invariant features. In: Proceedings of the 7th IEEE International Conference on Computer Vision. Kerkyra, Greece: IEEE, 1999. 1150−1157

8 Dalal N, Triggs B. Histograms of oriented gradients for human detection. In: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05). San Diego, USA: IEEE, 2005. 886−893

9 Qi Mei-Bin, Tan Sheng-Shun, Wang Yun-Xia, Liu Hao, Jiang Jian-Guo. Multi-feature subspace and kernel learning for person re-identification. *Acta Automatica Sinica*, 2016, **42**(2): 229−308 (齐美彬, 檀胜顺, 王运侠, 刘皓, 蒋建国. 基于多特征子空间与核学习的行人再识别. 自动化学报, 2016, **42**(2): 229−308)

10 Liao S C, Hu Y, Zhu X Y, Li S Z. Person re-identification by local maximal occurrence representation and metric learning. In: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, USA: IEEE, 2015. 2197−2206

11 Köstinger M, Hirzer M, Wohlhart P, Both P M, Bischof H. Large scale metric learning from equivalence constraints. In: Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition. Providence, USA: IEEE, 2012. 2288−2295

12 Huang T, Russell S. Object identification in a Bayesian context. In: Proceedings of the 15th International Joint Conference on Artifical Intelligence. San Francisco, USA: Morgan Kaufmann Publishers Inc., 1997. 1276−1282

13 Zajdel W, Zivkovic Z, Krose B J A. Keeping track of humans:

Have I seen this person before? In: Proceedings of the 2005 IEEE International Conference on Robotics and Automation. Barcelona, Spain: IEEE, 2005. 2081−2086

14 Gray D, Brennan S, Tao H. Evaluating appearance models for recognition, reacquisition, and tracking. In: Proceedings of the 10th IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS). Rio de Janeiro, Brazil: IEEE, 2007. 1−7

15 Li W, Zhao R, Xiao T, Wang X G. DeepReID: Deep filter pairing neural network for person re-identification. In: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, USA: IEEE, 2014. 152−159

16 Zheng Z D, Zheng L, Yang Y. Unlabeled samples generated by GAN improve the person re-identification baseline in vitro. In: Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV). Venice, Italy: IEEE, 2017. 3774-3782

17 Sun Y F, Zheng L, Yang Y, Tian Q, Wang S J. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In: Proceedings of the 15th European Conference on Computer Vision (ECCV 2018). Munich, Germany: Springer, 2018. 501−518

18 Yu H X, Zheng W S, Wu A C, Guo X W, Gong S G, Lai J H. Unsupervised person re-identification by soft multilabel learning. In: Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA: IEEE, 2019. 2143−2152

19 Wu A C, Zheng W S, Lai J H. Unsupervised person re-identification by camera-aware similarity consistency learning. In: Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea (South): IEEE, 2019. 6921−6930

20 Zheng X Y, Cao J W, Shen C H, You M Y. Self-training with progressive augmentation for unsupervised cross-domain person re-identification. In: Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea (South): IEEE, 2019. 8221−8230

21 Zheng L, Shen L Y, Tian L, Wang S J, Wang J D, Tian Q. Scalable person re-identification: A benchmark. In: Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV). Santiago, Chile: IEEE, 2015. 1116−1124

22 Ristani E, Solera F, Zou R, Cucchiara R, Tomasi C. Performance measures and a data set for multi-target, multi-camera tracking. In: Proceedings of the 2016 European Conference on Computer Vision. Amsterdam, The Netherlands: Springer, 2016. 17−35

23 Wei L H, Zhang S L, Gao W, Tian Q. Person transfer GAN to bridge domain gap for person re-identification. In: Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018. 79−88

24 Zheng L, Bie Z, Sun Y F, Wang J D, Su C, Wang S J, et al. MARS: A video benchmark for large-scale person re-identification. In: Proceedings of the 14th European Conference on Computer Vision (ECCV 2016). Amsterdam, The Netherlands: Springer, 2016. 868−884

25 Wang T Q, Gong S G, Zhu X T, Wang S J. Person re-identification by video ranking. In: Proceedings of the 13th European Conference on Computer Vision (ECCV 2014). Zurich, Switzerland: Springer, 2014. 688−703

26 Hirzer M, Beleznai C, Roth P M, Bischof H. Person re-identification by descriptive and discriminative classification. In: Proceedings of the 17th Scandinavian Conference on Image Analysis. Ystad, Sweden: Springer, 2011. 91−102

27 Zheng W S, Li X, Xiang T, Liao S C, Lai J H, Gong S G. Partial person re-identification. In: Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV). Santiago, Chile: IEEE, 2015. 4678−4686

28 Zheng W S, Gong S G, Xiang T. Person re-identification by probabilistic relative distance comparison. In: Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2011). Colorado Springs, USA: IEEE, 2011. 649−656

29 Cheng D S, Cristani M, Stoppa M, Bazzani L, Murino V. Custom pictorial structures for re-identification. In: Proceedings of the 22nd British Machine Vision Conference. Dundee, UK: BMVA Press, 2011. 1−11

30 Cho Y J, Yoon K J. Improving person re-identification via pose-aware multi-shot matching. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, USA: IEEE, 2016. 1354−1362

31 Toshev A, Szegedy C. DeepPose: Human pose estimation via deep neural networks. In: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, USA: IEEE, 2014. 1653−1660

32 Xiao B, Wu H P, Wei Y C. Simple baselines for human pose estimation and tracking. In: Proceedings of the 15th European Conference on Computer Vision (ECCV 2018). Munich, Germany: Springer, 2018. 472−487

33 Newell A, Yang K Y, Deng J. Stacked hourglass networks for human pose estimation. In: Proceedings of the 14th European Conference on Computer Vision (ECCV 2016). Amsterdam, The Netherlands: Springer, 2016. 483−499

34 Zhao H Y, Tian M Q, Sun S Y, Shao J, Yan J J, Yi S, et al. Spindle net: Person re-identification with human body region guided feature decomposition and fusion. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA: IEEE, 2017. 907−915

35 Zheng L, Huang Y J, Lu H C, Yang Y. Pose-invariant embedding for deep person re-identification. *IEEE Transactions on Image Processing*, 2019, **28**(9): 4500−4509

36 Miao J X, Wu Y, Liu P, Ding Y H, Yang Y. Pose-guided feature alignment for occluded person re-identification. In: Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea (South): IEEE, 2019. 542−551

37 Goodfellow I J, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial nets. In: Proceedings of the 27th Conference on Neural Information Processing Systems. Montreal, Canada: MIT Press, 2014. 2672−2680

38 Liu J X, Ni B B, Yan Y C, Zhou P, Cheng S, Hu J G. Pose transferrable person re-identification. In: Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018. 4099−4108

39 Zhu Z, Huang T T, Shi B G, Yu M, Wang B F, Bai X. Progressive pose attention transfer for person image generation. In: Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA: IEEE, 2019. 2342−2351

40 Szegedy C, Liu W, Jia Y Q, Sermanet P, Reed S, Anguelov D, et al. Going deeper with convolutions. In: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, USA: IEEE, 2015. 1−9

41 He K M, Zhang X Y, Ren S Q, Sun J. Deep residual learning for image recognition. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, USA: IEEE, 2016. 770−778

42 Mirza M, Osindero S. Conditional generative adversarial nets. arXiv preprint arXiv: 1411.1784, 2014.

43 Wei S E, Ramakrishna V, Kanade T, Sheikh Y. Convolutional pose machines. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, USA: IEEE, 2016. 4724−4732

44 Cao Z, Simon T, Wei S E, Sheikh Y. Realtime multi-person 2D pose estimation using part affinity fields. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE, 2017. 1302−1310

45 Fang H S, Xie S Q, Tai Y W, Lu C W. RMPE: Regional multi-person pose estimation. In: Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV). Venice, Italy: IEEE, 2017. 2353−2362

46 Cao Z, Hidalgo G, Simon T, Wei S E, Sheikh Y. OpenPose: Realtime multi-person 2D pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021, **43**(1): 172−186

47 Wang G S, Yuan Y F, Chen X, Li J W, Zhou X. Learning discriminative features with multiple granularities for person re-identification. In: Proceedings of the 26th ACM International Conference on Multimedia. Seoul, Korea (South): ACM, 2018. 274−282

48 Cheng D, Gong Y H, Zhou S P, Wang J J, Zheng N N. Person re-identification by multi-channel parts-based CNN with improved triplet loss function. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, USA: IEEE, 2016. 1335−1344

49 Zheng F, Deng C, Sun X, Jiang X Y, Guo X W, Yu Z Q, et al. Pyramidal person re-identification via multi-loss dynamic training. In: Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA: IEEE, 2019. 8506−8514

50 Luo H, Jiang W, Zhang X, Fan X, Qian J J, Zhang C. AlignedReID++: Dynamically matching local information for person re-identification. *Pattern Recognition*, 2019, **94**: 53−61

51 Sun Y F, Xu Q, Li Y L, Zhang C, Li Y K, Wang S J, et al. Perceive where to focus: Learning visibility-aware part-level features for partial person re-identification. In: Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA: IEEE, 2019. 393−402

52 Fu Y, Wei Y C, Wang G S, Zhou Y Q, Shi H H, Uiuc U, et al. Self-Similarity grouping: A simple unsupervised cross domain adaptation approach for person re-identification. In: Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea (South): IEEE, 2019. 6111−6120

53 Schroff F, Kalenichenko D, Philbin J. FaceNet: A unified embedding for face recognition and clustering. In: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, USA: IEEE, 2015. 815−823

54 Sun X X, Zheng L. Dissecting person re-identification from the viewpoint of viewpoint. In: Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA: IEEE, 2019. 608−617

55 Bak S, Zaidenberg S, Boulay B, Brémond F. Improving person re-identification by viewpoint cues. In: Proceedings of the 11th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). Seoul, Korea (South): IEEE, 2014. 175−180

56 Bialkowski A, Denman S, Sridharan S, Fookes C, Lucey P. A database for person re-identification in multi-camera surveillance networks. In: Proceedings of the 2012 International Conference on Digital Image Computing Techniques and Applications (DICTA). Fremantle, Australia: IEEE, 2012. 1−8

57 Wu Z Y, Li Y, Radke R J. Viewpoint invariant human re-identification in camera networks using pose priors and subject-discriminative features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2015, **37**(5): 1095−1108

58 Li D W, Zhang Z, Chen X T, Huang K Q. A richly annotated pedestrian dataset for person retrieval in real surveillance scenarios. *IEEE Transactions on Image Processing*, 2018, **28**(4): 1575−1590

59 Sarfraz M S, Schumann A, Wang Y, Stiefelhagen R. Deep view-sensitive pedestrian attribute inference in an end-to-end model. In: Proceedings of the 2017 British Machine Vision Conference. London, UK: BMVA Press, 2017. 134.1−134.13

60 Sarfraz M S, Schumann A, Eberle A, Stiefelhagen R. A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking. In: Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018. 420−429

61 Zhao L M, Li X, Zhuang Y T, Wang J D. Deeply-learned part-aligned representations for person re-identification. In: Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV). Venice, Italy: IEEE, 2017. 3239−3248

62 Song C F, Huang Y, Ouyang W L, Wang L. Mask-guided con−trastive attention model for person re-identification. In: Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018. 1179-1188

63 Hu J, Shen L, Sun G. Squeeze-and-excitation networks. In: Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018. 7132−7141

64 Nair V, Hinton G E. Rectified linear units improve restricted Boltzmann machines. In: Proceedings of the 27th International Conference on International Conference on Machine Learning. Madison, USA: Omnipress, 2010. 807−814

65 Li W, Zhu X T, Gong S G. Harmonious attention network for person re-identification. In: Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018. 2285−2294

66 Chen T L, Ding S J, Xie J Y, Yuan Y, Chen W Y, Yang Y, et al. ABD-Net: Attentive but diverse person re-identification. In: Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea (South): IEEE, 2019. 8350−8360

67 Fu J L, Zheng H L, Mei T. Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA: IEEE, 2017. 4476−4484

68 Dai Z X, Chen M Q, Gu X D, Zhu S Y, Tan P. Batch DropBlock network for person re-identification and beyond. In: Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea (South): IEEE, 2019. 3690−3700

69 Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. In: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, USA: IEEE, 2015. 3431−3440

70 Ghiasi G, Lin T Y, Le Q V. DropBlock: A regularization method for convolutional networks. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems. Montreal, Canada: Curran Associates Inc., 2018. 10750−10760

71 Zeiler M D, Fergus R. Visualizing and understanding convolutional networks. In: Proceedings of the 13th European Conference on Computer Vision (ECCV 2014). Zurich, Switzerland: Springer, 2014. 818−833

72 Zhou B L, Khosla A, Lapedriza A, Oliva A, Torralba A. Learning deep features for discriminative localization. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, USA: IEEE, 2016. 2921−2929

73 Peng P X, Xiang T, Wang Y W, Pontil M, Gong S G, Huang T J, et al. Unsupervised cross-dataset transfer learning for person re-identification. In: Proceedings of the 2016 IEEE Conference

on Computer Vision and Pattern Recognition (CVPR). Las Vegas, USA: IEEE, 2016. 1306−1315

74 Deng W J, Zheng L, Ye Q X, Kang G L, Yang Y, Jiao J B. Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. In: Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018. 994−1003

75 Guo J Y, Yuan Y H, Huang L, Zhang C, Yao J G, Han K. Beyond human parts: Dual part-aligned representations for person re-identification. In: Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea (South): IEEE, 2019. 3641−3650

76 Selvaraju R R, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV). Venice, Italy: IEEE, 2017. 618−626

77 Wu Yan-Cheng, Chen Hong-Chang, Li Shao-Mei, Gao Chao. Person re-identification using attribute priori distribution. *Acta Automatica Sinica*, 2019, **45**(5): 953−964
(吴彦丞, 陈鸿昶, 李邵梅, 高超. 基于行人属性先验分布的行人再识别. 自动化学报, 2019, **45**(5): 953−964)

78 Yang Q Z, Wu A C, Zheng W S. Person re-identification by contour sketch under moderate clothing change. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021, **43**(6): 2029−2046

79 Zhang L B, Huang S L, Liu W, Tao D C. Learning a mixture of granularity-specific experts for fine-grained categorization. In: Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea (South): IEEE, 2019. 8330−8339

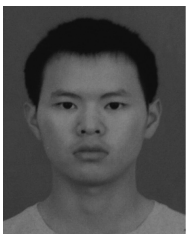
80 Simon M, Rodner E. Neural activation constellations: Unsupervised part model discovery with convolutional networks. In: Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV). Santiago, Chile: IEEE, 2015. 1143−1151

81 Xiao T J, Xu Y C, Yang K Y, Zhang J X, Peng Y X, Zhang Z. The application of two-level attention models in deep convolutional neural network for fine-grained image classification. In: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA: IEEE, 2015. 842−850

82 Yin J H, Wu A C, Zheng W S. Fine-grained person re-identification. *International Journal of Computer Vision*, 2020, **128**(6): 1654−1672

姚 足 西南交通大学计算机与人工智能学院硕士研究生. 主要研究方向为行人重识别和深度学习.
E-mail: yaozu@my.swjtu.edu.cn
(**YAO Zu** Master student at the School of Computing and Artificial Intelligence, Southwest Jiaotong University. His research interest covers person re-identification and deep learning.)

龚 勋 西南交通大学计算机与人工智能学院教授. 主要研究方向为图像处理, 模式识别及深度学习. 本文通信作者.
E-mail: gongxun@swjtu.edu.cn
(**GONG Xun** Professor at the School of Computing and Artificial Intelligence, Southwest Jiaotong University. His research interest covers medical image processing, pattern recognition, and deep learning. Corresponding author of this paper.)



陈 锐 西南交通大学计算机与人工智能学院硕士研究生. 主要研究方向为人脸识别和深度学习.
E-mail: richard3chen@gmail.com
(**CHEN Rui** Master student at the School of Computing and Artificial Intelligence, Southwest Jiaotong University. His research interest covers face recognition and deep learning.)

卢 奇 西南交通大学计算机与人工智能学院硕士研究生. 主要研究方向为人脸识别和深度学习.
E-mail: luqi@my.swjtu.edu.cn
(**LU Qi** Master student at the School of Computing and Artificial Intelligence, Southwest Jiaotong University. His research interest covers face recognition and deep learning.)

罗 彬 西南交通大学计算机与人工智能学院硕士研究生. 主要研究方向为行人重识别和深度学习.
E-mail: ansvic@icloud.com
(**LUO Bin** Master student at the School of Computing and Artificial Intelligence, Southwest Jiaotong University. His research interest covers person re-identification and deep learning.)