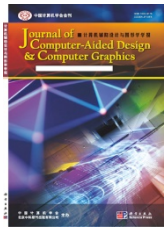

计算机辅助设计与图形学学报
Journal of Computer-Aided Design & Computer Graphics
ISSN 1003-9775,CN 11-2925/TP

《计算机辅助设计与图形学学报》网络首发论文

题目：监控场景下基于单帧与视频数据的行人属性识别方法综述及展望
作者：曹雨然，逯伟卿，于金佐，周亦博，胡海苗
收稿日期：2023-06-20
网络首发日期：2024-01-13
引用格式：曹雨然，逯伟卿，于金佐，周亦博，胡海苗．监控场景下基于单帧与视频数据的行人属性识别方法综述及展望[J/OL]．计算机辅助设计与图形学学报．https://link.cnki.net/urlid/11.2925.TP.20240112.1314.002

监控场景下基于单帧与视频数据的行人属性识别方法综述及展望

曹雨然[1,2)], 逯伟卿[1,2)], 于金佐[1,2)], 周亦博[1,2)], 胡海苗[1,2)*]

[1)] (北京航空航天大学虚拟现实技术与系统国家重点实验室　北京　100191)
[2)] (北京航空航天大学杭州创新研究院　杭州　310052)
(hu@buaa.edu.cn)

摘　要: 行人属性识别旨在判断目标行人的预定义属性标签，从而生成关于该行人的结构化描述，包括年龄、性别、衣着、配饰等多种层次的语义信息. 由于行人属性识别在视频监控领域具有极大的应用潜力，该任务广受研究者关注. 随着深度学习的快速发展，研究者提出众多识别行人属性的方法，以获得更为精准的识别结果. 针对当前复杂场景下，该任务面临的监控画面不清晰、行人状态变化、遮挡等问题，对监控场景下基于单帧与视频数据的行人属性识别方法进行综述，首先围绕行人属性识别这一任务，介绍其研究背景及任务概念，指出当前研究所面临的问题与挑战；其次根据“单帧图像”和基于视频数据的“序列图像”2 种不同的样本类型，对行人属性识别方法进行分类，并依据属性识别过程中所采用的技巧和思路，归纳总结最新提出的行人属性识别方法，概述研究现状；再对当前主流使用的数据集进行分析比较，总结其特点；最后，从状态引导行人属性识别、立体属性、多任务融合、新数据集构建 4 个方面，思考该领域的未来发展方向并作出展望.

关键词: 深度学习；智能视频监控；多标签分类；行人属性识别；数据集分析
中图法分类号: TP391.41　　**DOI:** 10.3724/SP.J.1089.2023.2023-00362

Pedestrian Attribute Recognition in Surveillance Scenario: A Survey and Future Perspectives on Frame vs. Video Based Methods

Cao Yuran[1,2)], Lu Weiqing[1,2)], Yu Jinzuo[1,2)], Zhou Yibo[1,2)], and Hu Haimiao[1,2)*]

[1)] (*State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing*　100191)
[2)] (*Hangzhou Innovation Institute, Beihang University, Hangzhou*　310052)

Abstract: Pedestrian attribute recognition aims to predict the predefined attributes of a target pedestrian, generating a structured description of the pedestrian, which includes levels of semantic information like age, gender, clothing, accessories and other levels of semantic information. Due to its wide application in the field of video surveillance and security, pedestrian attribute recognition has been widely concerned by researchers. With the rapid development of deep learning, researchers have proposed many methods to recognize pedestrian attributes in order to obtain more accurate results. In view of the challenges faced by this task in complex scenes, such as unclear surveillance scenes, pedestrian status change, occlusion, etc., this paper reviews image-based and video-based pedestrian attribute recognition methods in surveillance scenario. First, the research background and the concept of pedestrian attribute recognition are introduced, and the problems and challenges faced by the current research are pointed out. The pedestrian attribute recognition methods are classified ac-

收稿日期: 2023-06-20; 修回日期: 2023-12-11. **基金项目**: 国家自然科学基金(62122011, U21A20514), 浙江省“尖兵”研发攻关计划项目(2023C01030). **曹雨然**(2000—), 女，硕士研究生, CCF 学生会员，主要研究方向为深度学习、计算机视觉和模式识别；**逯伟卿**(1998—), 男，硕士，主要研究方向为深度学习、计算机视觉和模式识别；**于金佐**(1999—), 男，硕士研究生，主要研究方向为深度学习、计算机视觉和模式识别；**周亦博**(1995—), 男，博士研究生，主要研究方向为深度学习、计算机视觉和模式识别；**胡海苗**(1983—), 男，博士，教授，博士生导师, CCF 会员，论文通信作者，主要研究方向为计算机视觉、智能感知、视频分析与理解.

cording to two different sample types of “single frame” and “sequential frames captured from video”. The newly proposed methods are summarized on the basis of skills and ideas adopted in the attribute recognition process, overviewing the research status. Then the current commonly employed datasets and experimental results are analyzed. Finally, from the four aspects of state-guided pedestrian attribute recognition, tri-dimensional attribute, multi-task fusion and new data set construction, the future direction of this field is prospected.

Key words: deep learning; intelligent visual surveillance; multi-label classification; pedestrian attribute recognition; datasets analysis

安防业务需求的日益增长，使视频监控系统广泛应用于安保、刑侦和交通等诸多领域. 随着视频监控在世界范围内的普及，人工检索的方式越来越难以处理监控摄像头采集到的庞大数据. 在此背景下，Zhu 等[1]首次提出行人属性识别(pedestrian attribute recognition, PAR)这一视觉任务，通过对行人的性别、年龄、服装、配饰属性的智能识别建立行人结构化描述，为实现自动化的监控行人检索提供强有力的支撑，使得图像和视频内容的智能识别成为可能.

行人属性识别任务中，一般以行人检测框内的图像作为输入，获取行人目标，将预定义列表中的每个属性作为分类任务，通过研究者设计的模型进行处理；以属性识别结果作为输出，将非结构化的图像数据转化为一系列属性，生成该行人的结构化描述.

随着深度学习在计算机视觉领域的发展，研究者们提出一系列提升行人属性识别效果的方法. Wang 等[2]将 2014—2020 年的行人属性识别方法分为 8 类：基于行人全局图像的方法、基于行人局部图像的方法、基于注意力机制的方法、基于时序预测的方法、设计新损失函数的方法、基于课程学习的方法、基于图模型的方法和其他算法，并从设计更准确有效的定位算法、使用模型进行数据增强、进一步发挥注意力机制的效能等角度，对该领域未来的发展方向进行分析.

Yaghoubi 等[3]梳理了行人属性识别任务需解决的 5 类挑战：数据分布不平衡、数据的局限性、属性间关系的不充分利用、无法对属性相关区域进行准确定位和遮挡问题，并提出以挑战为导向的方法分类方式，将现有工作按照其算法设计动机应对的上述某类挑战进行归类，并分析总结. 本文未对人体属性识别的未来发展方向展开详细阐述.

贾健等[4]从行人属性识别具体应用场景的层面，对监控场景中的行人属性识别研究进行梳理概括，从技术侧重点将“监控场景中”的行人属性识别和“面向行人再识别”的行人属性识别进行区分，以时间为脉络，重点概括分析基于深度学习的行人属性识别方法，并针对该领域的解耦问题、属性关系模型和跨域问题3项挑战探讨未来发展方向.

不同于以上综述文献，本文主要对近 3 年来新提出的行人属性识别方法进行归纳概述，将该领域目前面临的挑战分为现实问题和数据集问题 2 个方向进行分析；再根据单帧图像和多帧序列图像 2 种数据类型，将行人属性识别分为 2 类方法，并根据方法机制对基于单帧图像的识别方法分为定位属性相关区域、改进特征提取机制挖掘属性关系 2 类方法，将基于序列图像的方法分为融合时间注意力机制和改进注意力机制及损失函数两类；再介绍主流使用的数据集以及相关方法在数据集上的实验结果，最后从利用状态引导行人属性识别、构建立体属性、多任务学习和构建新数据集 4 个方面对行人属性识别的未来发展方向作出展望.

1 问题与挑战

用于行人属性识别的图像大多采集自监控摄像头，会受到因拍摄距离较远与摄像头配置较低等因素导致的成像分辨率不足的影响；同时，行人属性识别在实际应用中面临着自然环境条件变化、行人姿态变化、环境及其他行人遮挡等复杂场景，而用于行人属性识别任务的数据集通常存在场景限定导致数据偏置、属性分布不均衡等现象. 因此，行人属性识别任务主要面临现实问题以及数据集问题 2 个方面带来的挑战.

1.1 现实问题

1.1.1 图像分辨率低

用于行人属性识别任务的输入图像一般来源

于未进行清晰化处理的监控录像，通常目标行人在画面中所占比例较小，加之监控摄像头与行人之间距离往往较远，如图 1 所示，在有限的摄像头配置下易导致图像拍摄清晰度不足，为细粒度属性的识别带来了困难.

图 1　部分行人图像分辨率较低

1.1.2　自然环境因素变化

无论是室内监控或是室外监控，摄像头都处于全天候工作状态，一天之中时间的变化也会使得拍摄环境的光线产生变化. 对室外监控而言，天气的变化也会带来光线的变化. 如图 2 所示，光线不足使得画面色彩较暗且画质不够清晰，而过于强烈的光线则会导致过曝现象的发生. 因此，光线变化会对色彩等属性的识别产生负面影响.

图 2　环境光线强弱变化

天气因素也会对室外监控摄像头采集到的图片质量产生严重影响. 如图 3 所示，监控画面易受到雨雾天气带来的成像干扰，需要辅之以除雾等图像清晰化技术[5].

图 3　去雾处理示例

1.1.3　行人状态变化

如图 4 所示，在监控录像中，大多数行人都处于持续的运动状态，其位置、姿态随时间产生变化. 因此，从监控录像中截取到的某一帧画面里的行人可能正处于移动状态，致使运动模糊，增加属性识别难度. 同时，摄像头与行人之间并不固定的相对位置和运动状态易导致多样且无法预测的行人成像视角，如图 5 所示，行人的朝向和俯仰角度并不固定. 说明模型需要理解不同成像视角带来的人体拓扑结构上的改变，以实现鲁棒的属性检测性能.

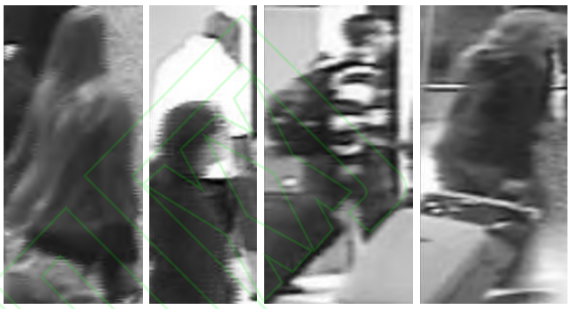

图 4　运动状态下的行人图像较为模糊

图 5　行人朝向变化及视角变化

1.1.4　遮挡现象

受制于行人的运动状态、姿态变化、朝向变化及其所处的环境，监控视角下的行人不可避免地会被自身肢体或物品、其他行人等遮挡部分肢体，使得部分属性处于不可视区域而难以准确识别，常见的遮挡情况示例如图 6 所示.

a. 行人与行人之间遮挡　　b. 行人自身或环境物品遮挡

图 6　遮挡示例

1.2　数据集问题

1.2.1　场景限定性

同一数据集中的行人图像大多采集自同一场景或相似场景，如图 7 所示，Market-1501-attribute

数据集中的行人图像均采集于夏季室外，行人的服装属性因季节限定较为单一，集中表现为短袖、短裤、裙装等夏装；RAP(richly annotated pedestrian)数据集中的行人图像均采集于室内商场，光线条件和场景背景缺乏多样性；不同数据集对行人标注的属性也不尽相同. 由此可以看出，单个数据集中的行人图像仅能反映这一限定的应用场景下的行人属性特征，很难体现其他场景下的特征，导致当前属性识别方法的泛化性能较弱.

图 7 Market-1501-attribute 数据集样本示例

1.2.2 属性分布不平衡

属性分布不平衡是机器学习任务中的常见问题，行人属性识别任务也不例外. 数据集中的行人大多分布在青年及中年这 2 个年龄阶段，而儿童和老年人则占少数；此外，摄像头下采集到的行人角度主要为正向和背向，其比例明显高于侧向样本. Deng 等[6]指出，对于二元属性，2 个分类中的样本数量之比不应超过 20∶1，否则将被视为分布不平衡. 根据此条标准，PETA(pedestrian attribute)数据集中近一半的二元属性存在分布不平衡的问题，部分属性分布情况如图 8 所示. 样本数量过小使得训练效果不理想，而 2 个分类占比差距过大也会对训练结果造成负面影响.

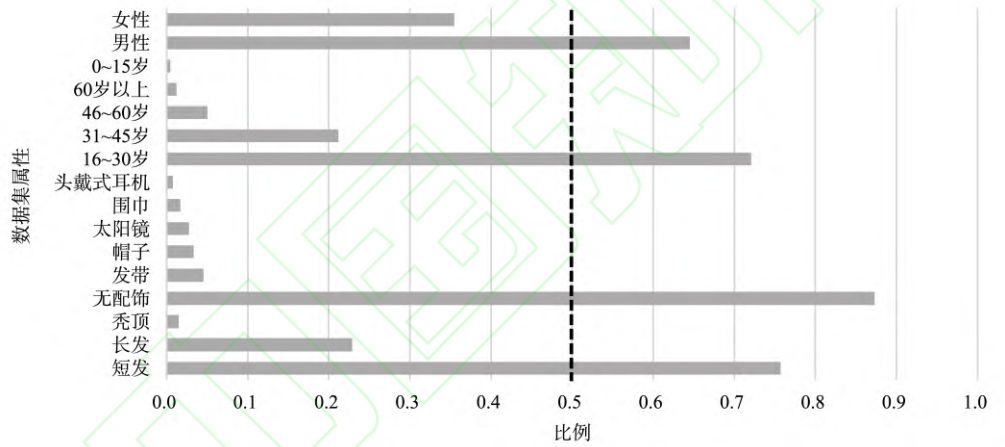

图 8 PETA 数据集部分属性比例分布情况

1.2.3 数据泄漏

在 PETA，RAP 等数据集中，训练集和测试集的划分标准较为随机，导致同一行人的图像样本可能会在训练集和测试集中同时出现. 如图 9a 为训练集样本示例，图 9b 为测试集样本示例，其中，同一线型方框中的样本表示在训练集和测试集中重复出现的行人，该现象会对属性识别的结果造成负面影响，不利于提升模型的泛化能力.

a. 训练集样本

b. 测试集样本

图 9 PETA 数据集中训练集及测试集样本示例

1.2.4 属性标注不完善

当前，大多数数据集将二元属性以“0”和“1”的形式进行标注，0 表示不存在，1 表示存在，这一标注方式混淆了“属性存在但处于不可视区域”和“属性不存在”2 个概念. 如图 10 所示，以属性“手提包袋”为例，图 10a，图 10b 两组的左图中的行人均有一只手处于不可见区域，而从每组的右图可以判断，图 10a 中的手提包袋为存在但处于不可见区域；

图 10b 中该属性为不存在.

a. 属性不可见　　b. 属性不存在

图 10　2 种属性对比

2　研究现状分析

处理行人属性识别任务的基本思路是将其视作多标签分类任务[7-9]. 若某个属性的判断结果有 2 种情况，则每个属性的识别过程为一个二分类问题；若某个属性的判断结果有多种情况，则将每种情况的判断视作一个二分类问题，再根据每种情况判断结果的置信度确定该属性的最终识别结果.

行人属性识别任务的定义可表示为如下形式：将由行人图像 x_i 及其相应属性标签 y_i 组成的数据集记作 $D=\{(x_i,y_i),i=1,2,3,4,\cdots,N\}$，其中，$N$ 表示行人样本数量，(x_i,y_i) 采样于某一联合概率分布 $P_{X\times Y}$. 在行人属性识别模型在 D 上完成训练后，输入任意采样于 P_X 的样本 x，期望得到对应的属性标签 $\tilde{y}=\arg\max_y Q_\theta(x,y)$，其中，$Q_\theta$ 表示属性识别模型对 $P_{X\times Y}$ 的估计，θ 表示该模型的参数.

神经网络的诞生使基于深度学习的行人属性识别方法逐渐成为主流. 研究者们基于卷积神经网络(convolutional neural network, CNN)、图卷积网络(graph convolutional network, GCN)和循环神经网络(recurrent neural network, RNN)等提出了众多行人属性识别方法，并取得了显著的进展. Li 等[8]提出 DeepMAR 网络，首次将深度学习引入行人属性识别领域，并利用多属性联合学习框架和加权损失函数识别行人属性；Li 等[10]首次将图卷积应用于行人属性识别，提出 VSGR(visual-semantic graph reasoning)方法，用于挖掘区域之间的空间关系和属性之间的语义关系；Wu 等[11]采用 RNN，借助自注意力模块提出序列上下文关系学习模型，构建空间和语义之间的关系模型.

CNN 可以更有效地提取属性的细粒度特征，GCN 可被用于挖掘属性的空间关系以及语义联系，而 RNN 为处理图像序列提供了有力的工具；结合关键点检测、新设计的损失函数等模块，深度学习极大地推动了行人属性识别的发展.

如图 11，本文依据单帧图像和序列图像 2 种输入数据类型，对行人属性识别方法进行分类，将基于单帧图像的行人属性识别方法根据其方法机制分为定位属性相关区域、改进特征提取机制挖掘属性关系 2 个方向，将基于序列图像的方法分为融合时间注意力机制和改进注意力机制及损失函数两类. 各行人属性识别方法概览如表 1 所示.

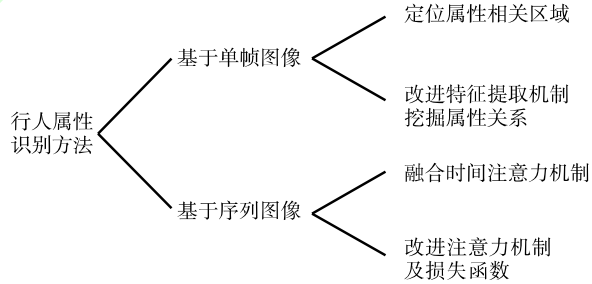

图 11　行人属性识别方法分类思路

表 1　行人属性识别方法概览

方法 (出版年)	骨干网络	单帧 图像	序列 图像	空间关系 /注意力	时间 注意力	属性 关系	上下文 关系	特点	局限性
P-Net (2019)	GoogleNet	√		√				根据相应的身体区域对属性进行分类，采用像素级分类判断属性相关区域	
DTM+AWK (2020)	ResNet-50	√		√				加入人体关键体模块作为辅助，无需增加额外计算	对低分辨率的图片易出现人体姿态的误判
SSC	ResNet-50	√		√			√	构建空间注意力和属性语义	

(2021)						联系 2 个板块，定位属性相关的空间区域，挖掘同一属性在不同图片之间的关系	
Semantic parsing-PAR (2021)	EfficientNet	√	√		√	利用语义解析分解行人图片，定位于属性相关的区域；利用特征金字塔结构挖掘属性关系	
HR-Net (2021)	GoogLeNet ResNet-50	√	√			将属性按照语义层次和抽象程度分为 3 个等级，利用属性之间低到高推测+高到低引导的双向关系辅助属性识别	
DAFL (2022)	ResNet-50	√	√			利用空间注意力对不同属性挖掘不同的特征表达	
JRL (2017)	AlexNet	√	√	√	√	同时挖掘属性之间的关系及属性与上下文环境间的关系，转化为序列预测问题	按照序列顺序识别属性，时间成本较高
JLAC (2020)	ResNet-50	√		√	√	利用 GCN 挖掘属性关系和上下文关系	
HFE 框架 (2020)	ResNet-50	√		√		构建属性和行人身份 2 个级别的特征嵌入，使用绝对边界正则化和动态损失权重设计新的损失函数	
CGCN (2022)	ResNet-101	√	√	√		同时挖掘属性之间的显性和隐性 2 类关系	
Label2Label (2022)	ResNet-50	√		√		将“属性”看作“词语”，利用语言模型框架进行属性识别	
OAGCN (2023)	swin Transformer	√	√	√		引入行人状态中的“朝向”，引导行人属性相关区域定位和关系学习	
时间注意力策略模型 (2019)	ResNet-50	√	√	√		首个基于视频的识别方法；引入时间注意力模块，将属性分为行为和身份相关 2 类	独立识别各属性，忽略属性之间的联系
STAM (2021)	ResNet-50	√	√	√		利用时间注意力引导模型关注无遮挡帧以应对遮挡问题	
MTA-Net (2020)	ResNet-152	√		√		挖掘利用之前、当下和之后 3 种时间状态下的上下文信息，引入焦平衡损失函数处理属性分布不平衡的问题	

2.1 基于单帧图像的行人属性识别方法

早期，研究者通常以整幅行人图像作为输入，这类方法导致对细粒度属性有价值的部分信息被忽略. 为了解决这一问题，部分研究者提出先对行人图像进行分割，再将局部图像作为输入的思路. 为了更准确地分割行人图像、定位与特定属性相关的局部区域，研究者加入行人关键点定位检测、注意力机制等辅助模块；此后，通过挖掘属性间的关系和上下文关系等信息、改进特征提取机制，进一步提升属性识别准确性，于是更多的处理技巧被运用到该任务中.

2.1.1 定位属性相关区域

通常，部分属性的识别与其空间位置之间存在紧密联系. 例如，对于“T 恤”“长袖”属性的判断，只需关注行人的上半身；对于“运动鞋”“皮鞋”属性的判断，只需关注行人的足部. 因此，研究者们通过分割图像、引入人体关键点等先验知识进行引导，对行人属性进行定位，以识别局部区域的细粒度属性.

P-Net(part-guided network). 由于某些属性与位置有着较强的关系，如与帽子、眼镜相关的属性通常需要关注行人的头部进行判断，而与服饰、背

包有关的属性则需关注行人的上身来进行判断. 因此, Bourdev 等[12], Zhang 等[13], Zhu 等[14], Zhou 等[15]和 Zheng 等[16]提出将图像分割为多个身体部位，利用位置约束与属性之间的关系识别行人属性的方法；但这些方法将原先的端到端分类任务转化为两阶段的分段任务，即先将图像分割成不同的部分，再识别行人属性，增加了计算成本.

An 等[17]提出一种新的引导网络 P-Net，将属性与身体的相应部分关联起来，其结构如图 12 所示. P-Net 将行人属性归为 6 个与身体部位相关的类：头部区域相关属性、上身相关属性、附件相关属性、下身相关属性、足部相关属性和全身相关属性. P-Net 的处理过程由 2 个模块组成：第 1 个模块通过 GoogLeNet 提取行人图像的特征；第 2 个模块则用于引导注意力，得到属性属于与身体部分相关的 6 类之中某一类的概率，从而生成注意图，在识别某一属性时更关注其所对应的属性区域，获取最终识别结果.

图 12　P-Net 的结构[17]

DTM-AWK(deep template matching-attribute-wise keypoints)方法. 定位属性的相关区域可以更准确地预测部分属性. 如图 13 所示，Zhang 等[18]提出一种基于 DTM 的方法来提取局部特征，并增加了 AWK 模块辅助属性识别，借助该模块引导属性相关区域的定位，利用人体姿态关键点作为先验知识监督 DTM 学习，该方法不需要额外的计算复杂度，节省了计算成本；但对于低分辨率的图像, DTM-AWK 方法可能会对人体姿态出现误判，应对这类输入图像时其效果不佳.

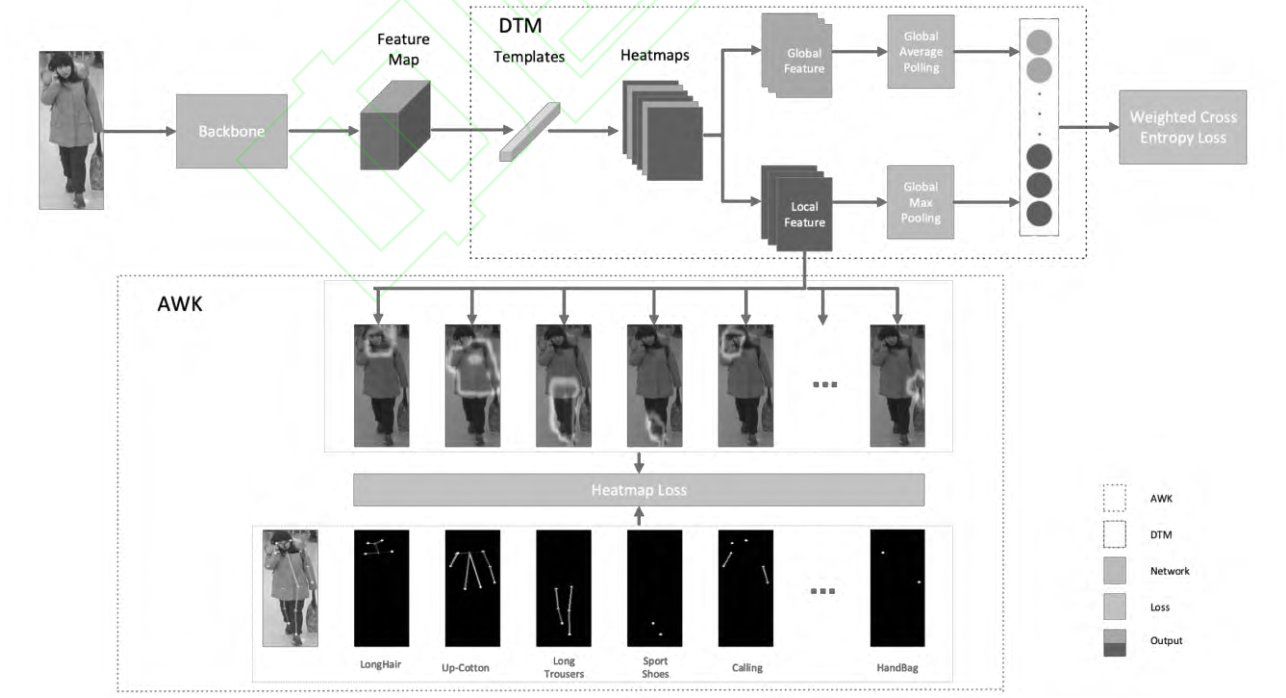

图 13　DTM-AWK 方法的结构[18]

SSC(spatial and semantic consistency)框架. Jia 等[19]提出 SSC 框架，如图 14 所示，该框架包含空间一致性(spatial consistency, SPAC)和语义一致性(semantic consistency, SEMC)这 2 个模块，以充分挖掘相同属性的属性相关信息，其中，SPAC 模块与空间一致性正则化相结合，致力于处理空间注意区域的偏差问题，生成与属性相关的可信空间区域，称为 SSC_{soft} 方法，SEMC 模块致力于从不同

样本中提取每个属性的图像间内在和判别性语义特征，并消除与属性无关的特征的干扰；此外，通过在数据集上进行实验来分析这 2 个模块的效果，并介绍了 SSC_{soft} 方法的 2 种变体 SSC_{hard} 和 SSC_{fix}，证明 SSC 的有效性. 在此之前研究属性关系的工作[7,20-22]主要关注单个图像，而 SSC 框架则同时考虑了多幅图像的空间和语义关系.

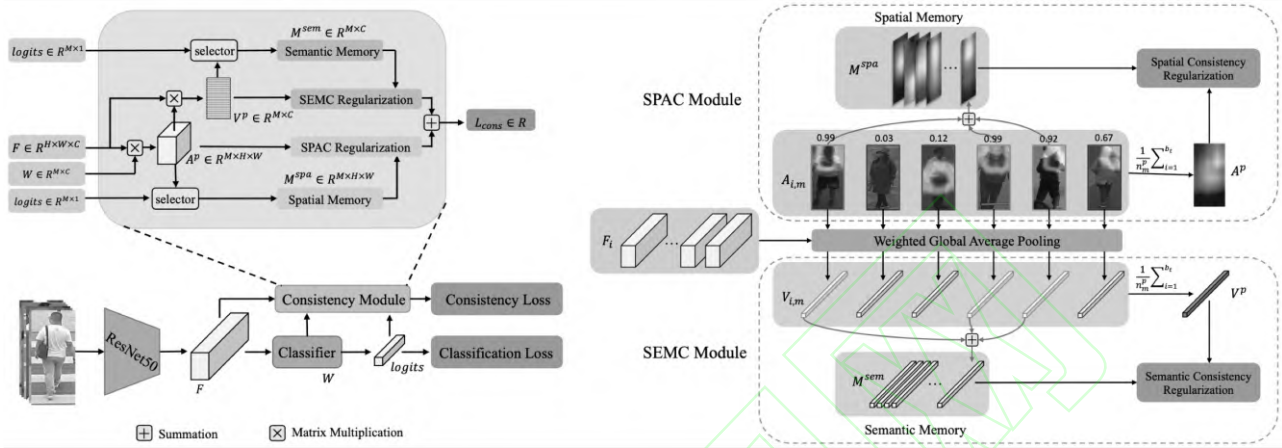

图 14 SSC 框架的结构[19]

Semantic Parsing-PAR. Moghaddam 等[23]将语义解析技术与行人属性识别相结合，同时利用特征金字塔结构挖掘上下文信息. 如图 15 所示，Semantic Parsing-PAR 共有行人语义分析和多尺度特征提取 2 个分支，用于联合挖掘语义和空间信息. 其中，通过 EfficientNet 实现特征提取，获取特征图谱；在此基础上，再利用语义分析提取人体部位，将行人图片分为前景、头部、上身、手、下身和鞋 6 个部分，实现在消除背景的情况下同时挖掘人体的语义信息和空间信息. 除了已有的图像级标注外，该方法还使用语义解析生成像素级注释，Moghaddam 等还提出属性识别-双向特征金字塔网络(attribute recognition-bidirectional feature pyramid network, AR-BiFPN)[24]，联合使用低级语义信息和高级语义信息来提高行人属性识别的准确性.

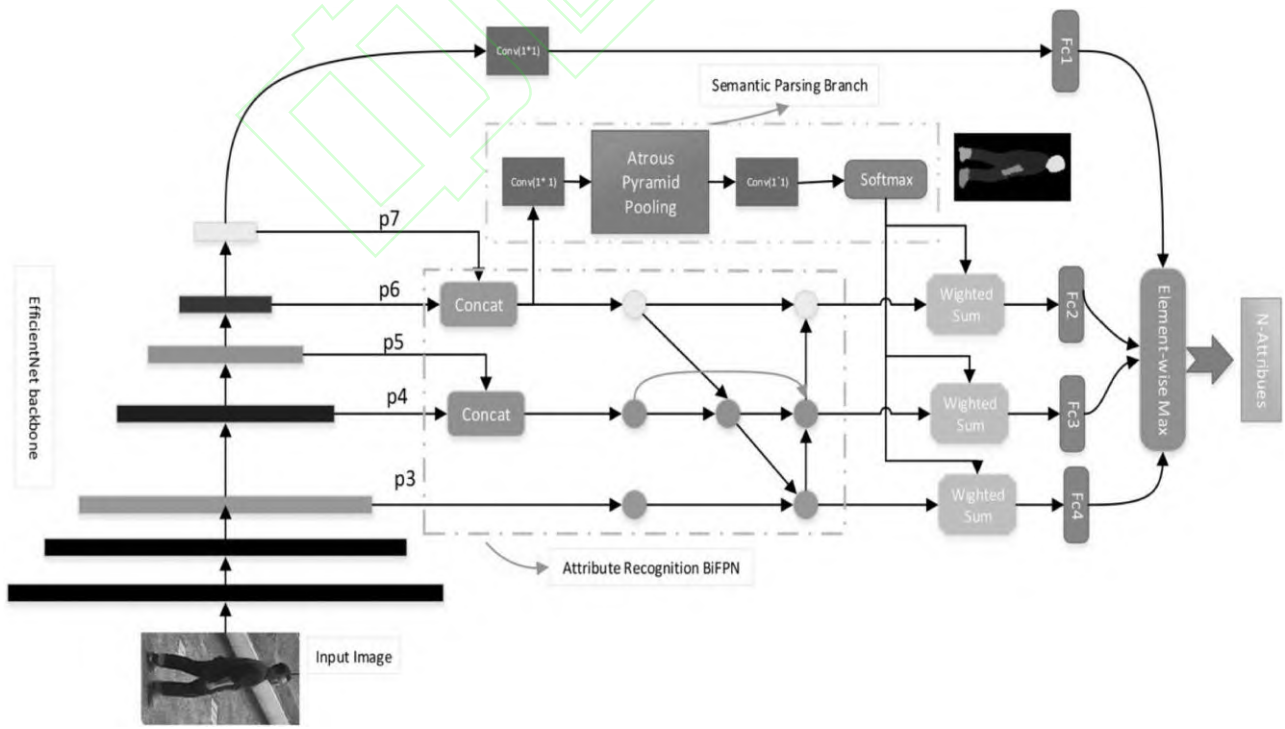

图 15 Semantic Parsing-PAR 方法的结构[23]

HR-Net(hierarchical reasoning net). 为了利用不同属性之间的双向关系，An 等[25]提出了 HR-Net，根据属性的语义级别和抽象程度，将属性分为三个层次类别：将条纹和格纹等高度依赖边缘和纹

理等低级信息的属性划分为低级属性；将性别、年龄等高级语义特征的属性划分为高级属性；服装、配饰等同时使用纹理信息和语义特征进行识别的属性划分为中级属性.

在 HR-Net 中，每个层次的识别被视为一个独立的学习任务，识别不同层次的属性时，HR-Net 会生成不同的特征图为属性识别提供信息. 与稀疏激活相比，HR-Net 的特征图将根据不同层级的属性关注人体的不同区域. 对于低级属性，该网络将提取更多的边缘及纹理信息；对于中级属性，同时提取语义信息及细节信息. 如图 16 所示，通过马尔可夫逻辑网络，较低层次的属性可以作为较高层次的属性的推理基础，以此提高高级属性的预测精度，且较高层次属性的判断结果也可引导较低层次的属性识别.

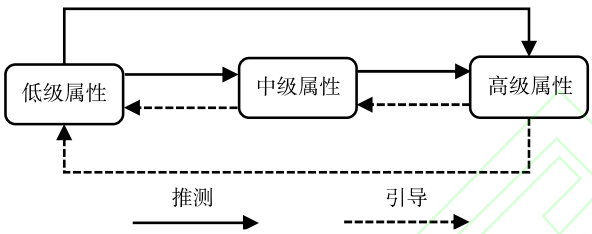

图 16　多层次属性之间的联系示意图

DAFL(disentangled attribute feature learning). Jia 等[26]指出，多个属性共享同一图像特征表达忽略了不同属性在空间和语义上的区别，不利于准确识别属性，由此提出 DAFL 框架，如图 17 所示，该框架由语义空间交叉注意力(semantic-spatial cross-attention, SSCA)模块和组注意力融合(group attention merging, GAM)模块组成. 其中，SSCA 模块用于提升属性相关区域的空间定位，并将区域特征聚合为多个解耦的属性特征；GAM 模块根据属性的空间分布对属性进行分组，利用可靠的组注意力监督查询注意力图的学习. DAFL 框架为每个属性生成独立的特征表达，充分利用不同属性的空间分布差异和语义特征差异提升识别效果.

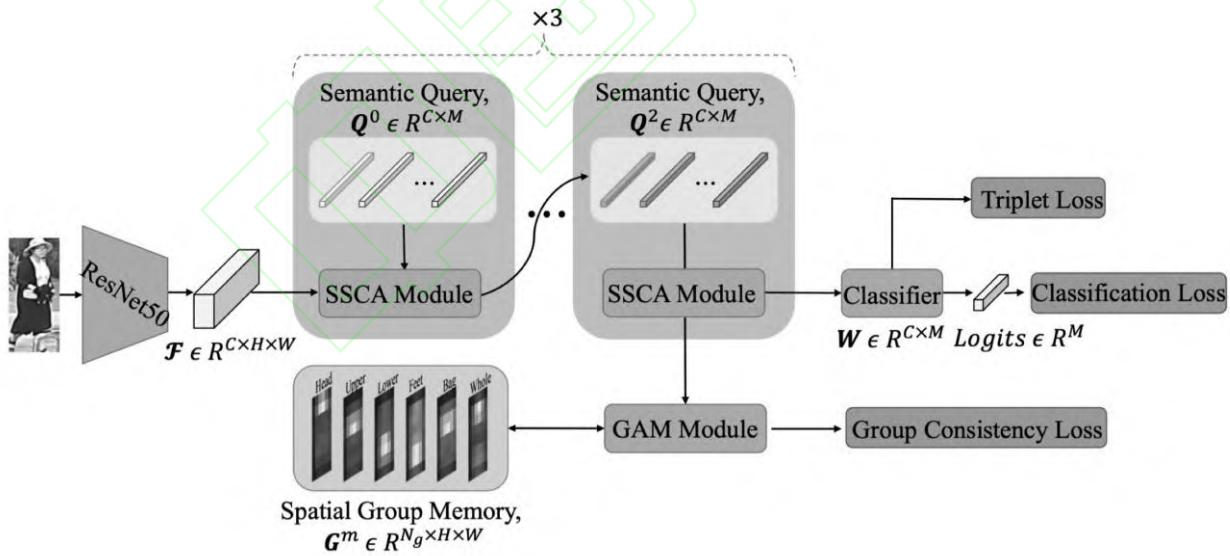

图 17　DAFL 框架的结构[26]

从人类认知的直观角度出发，定位属性相关区域有助于迅速识别属性；而采用属性空间关系的早期方法大多是通过先验知识人为地划分了属性相关区域，导致其对行人姿态变化以及视角的俯仰变化缺乏鲁棒性，且不能直接适用于仅拍摄到半身的行人图像.

空间注意力机制的运用，使得行人属性识别模型将更多的注意力集中在与属性识别相关的关键区域，增强了关键区域特征的重要性；同时，有选择性地关注与属性识别相关的信息，减少冗余信息和噪声的影响，规避非关键区域特征的干扰，提升了识别效果.

定位属性相关区域的方法存在的问题如下：目前数据集中的样本以平视角度采集为主，俯、仰角度的样本所占比例较小，对于“手提包”等与“手部”相关的属性通常被定位在图像中间两侧位置，

而实际中，当行人的姿态改变时，其手部位置也会随其动作的改变而改变；此外，俯、仰角度下行人的空间拓扑结构也不同于平视角度下的结构，视角的变化也是该领域需要解决的问题之一.

2.1.2 改进特征提取机制挖掘属性关系

早期的行人属性识别方法通常独立识别每个属性，忽略了属性之间的相关性. 事实上，这些行人属性可以根据其抽象程度及语义级别分为多层次，并且部分属性之间存在一定的相关性，包括正相关和负相关 2 种情况. 例如，“女性”和“裙子”更有可能同时存在，而“短裙”和“短裤”这 2 个属性通常不会同时出现. 此外，监控场景下采集到的行人图像往往存在画质模糊和行人被遮挡的问题，利用属性相关性进行合理推测，可以增强模糊区域以及被遮挡区域的属性的识别.

JRL(joint recurrent learning)模型. 由于监控画面普遍存在分辨率较低等问题，使得细粒度特征的识别准确率不够. 解决这类问题的思路通常有 2 类：一是挖掘属性之间的关系，如女性和短裙之间存在较强的联系，因此这 2 个属性很有可能同时存在；二是挖掘上下文环境信息作为识别的辅助，处于同一场景下的行人往往会具备相同的属性，如滑雪者更有可能佩戴护目镜. 在早期的方法中，这 2 种思路通常是独立研究的.

Wang 等[27]结合 RNN 和长短期记忆(long short term memory, LSTM)机制，提出 JRL 模型，不仅可用于探索属性之间的相关性，还可探索属性和上下文之间的关系. JRL 模型的结构如图 18 所示，通过对输入图像进行水平裁剪，将其变换为 6 个区域序列，并将预设列表中的属性按照随机或出现频率的顺序进行排列，再依次进行处理. Wang 等还引入一种数据驱动的注意机制用于特征提取，以更好地定位与属性相关的区域. 这是研究者们首次将行人属性识别任务视作为序列预测问题；但由于该模型中属性的识别是先后进行的而非同时处理，导致时间成本较高.

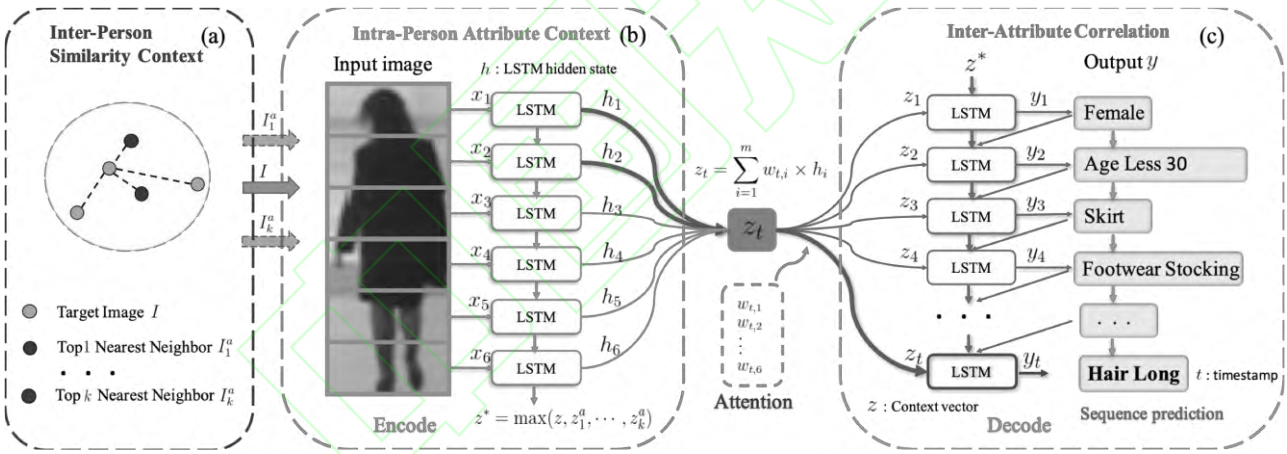

图 18 JRL 模型的网络结构图[27]

JLAC(joint learning of attribute and contextual relations). Tan 等[28]利用 GCN 提出一种新型的端到端网络，利用属性和上下文关系识别行人属性，称为 JLAC；其由 ARM(attribute relation module)和 CRM(contextual relation module)2 部分组合成一个双分支网络，同时学习 2 种关系. JLAC 的结构如图 19 所示，ARM 构造了一个具有特定属性特征的属性图，学习到的每个特征都被认为是图中的一个节点；CRM 引入一种图投影方案，将二维特征图投影到一组来自不同图像区域的节点上，将区域或像素的簇作为图的节点，再由该网络完成聚类任务，并在 2 个模块中分别使用 GCN 来分析多个属性之间的相关性和探索区域之间的上下文关系.

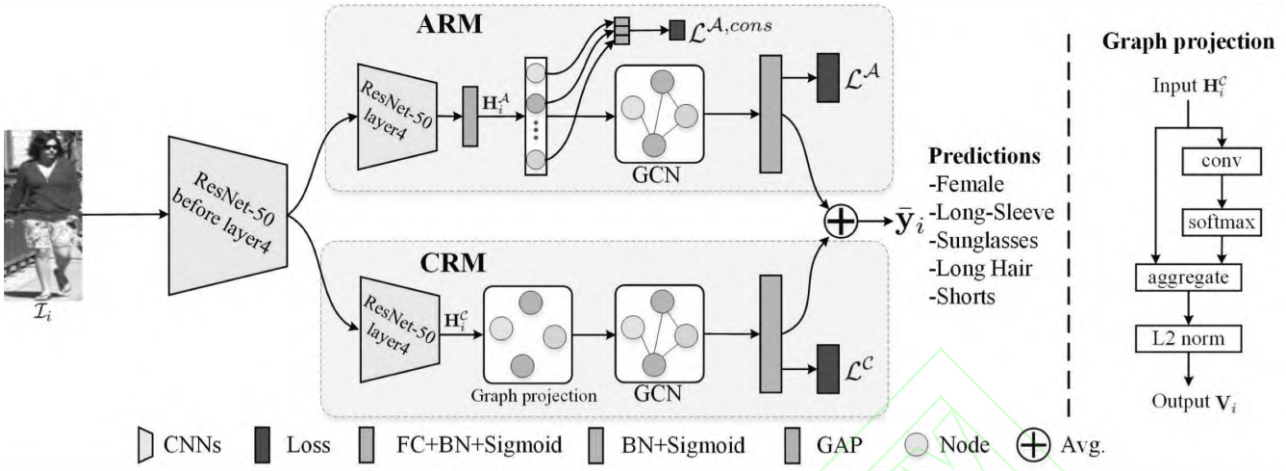

图 19　JLAC 方法的结构[28]

HFE(hierarchical feature embedding)框架. 在目前的属性分类中，同一属性的样本之间仍存在一些未被表述的类内差距. 例如，同为背包，不同行人的背包可能拥有不同的颜色特征和材质特征；同为帽子，不同行人可能戴着不同类别的帽子，如遮阳帽、毛线帽等；而对于同一行人，无论光线、视角是否变化，都可以认为该行人在不同帧中具有相同的属性.

基于来自同一行人身份的样本所具备的属性相同这一设想，Yang 等[29]提出一种用于属性识别的端到端 HFE 框架，可用于联合挖掘属性级别的特征及行人身份级别的特征. 如图 20 所示，HFE 框架可以基于属性级别，将具备同一属性特征的行人样本进行聚类，利用行人身份信息进行约束，在拥有相同属性的大类内将身份相同的样本进行聚类，使得具有相同属性的样本和具有相同行人身份的样本更为紧密. 该框架融合了属性和行人身份 2 个级别的特征，能够实现同时保留类内和类间 2 种特征嵌入. 在此基础上，Yang 等[29]还设计了类内和类间 2 个级别的 HFE 损失函数，并利用动态损失权重使其随着学习过程逐渐增加，再通过绝对边界正则化，使得类与类之间的边界更为显著.

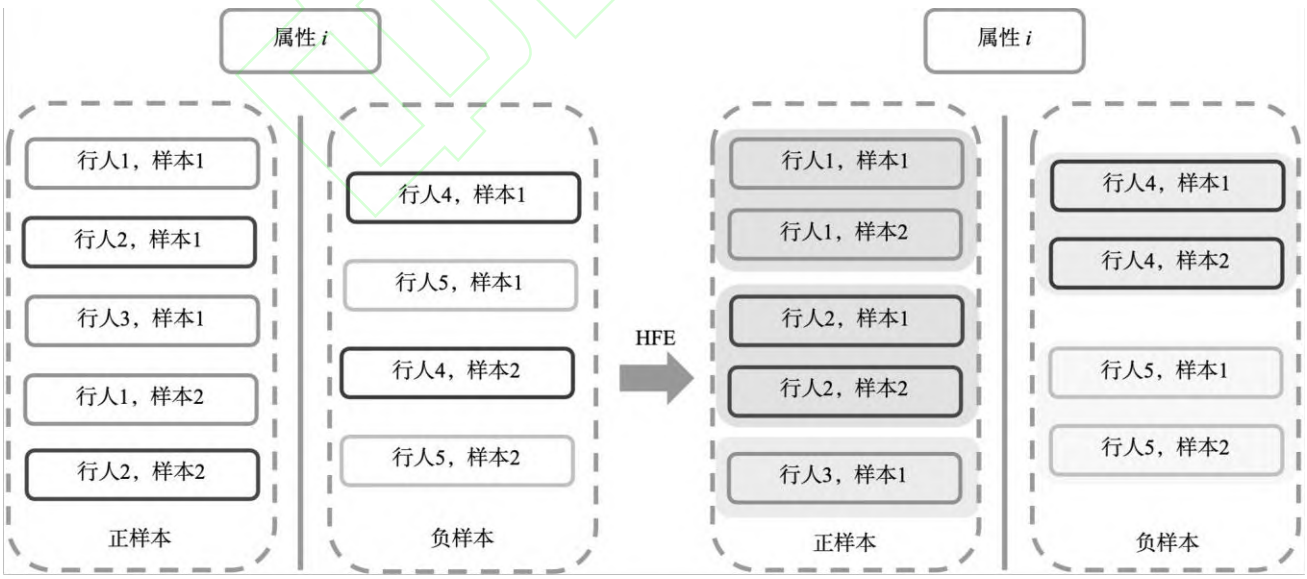

图 20　HFE 框架示意图

CGCN(correlation GCN). Fan 等[30]提出 CGCN 方法来充分挖掘属性之间的显性及隐性关系，如图 21 所示，其中，显性关系包括属性与空间位置间的关系，以及属性与属性之间的等级关系. 该方法通过建立特征向量将属性与图特征结合，利用自注意生成关系矩阵，通过图卷积实现属性关系的多层转移.

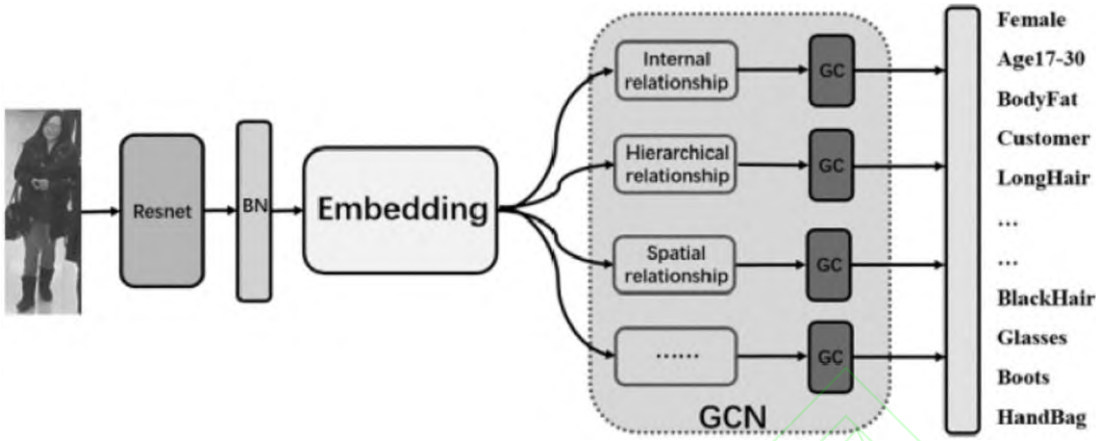

图 21 CGCN 方法的结构[30]

早期，与属性关系有关的研究大多只关注属性之间的单向关系，而 CGCN 方法的突出之处在于其引入了一个更全面的框架，由属性嵌入、关系矩阵构建和关系转移 3 个模块组成，充分利用了属性之间的双向关系、传递关系等更为复杂的联系，使其更具灵活性和拓展性.

Label2Label. Li 等[31]提出一种用于多属性学习的语言模型框架 Label2Label，并将其运用至人脸识别、行人属性识别、服饰属性识别等视觉任务中. Label2Label 框架示意图如图 22 所示，其中，图像的每个“属性”被视作一个“单词”，图像的所有属性视作由词语组成的无序但有意义的“句子”，通过这些词句挖掘属性之间的联系. 该框架由属性查询网络(attribute query network, AQN)和基于图像条件的掩码语言模型(image-conditioned masked language model, IC-MLM)组成. 其中, AQN 用于生成初始预测结果，并将这一结果作为 IC-MLM 的输入，通过随机掩蔽一些单词，并根据余下的句子和图像上下文推测这一缺失的单词，该模型利用此思路学习图像实例级别的属性关系，提升识别结果的准确率.

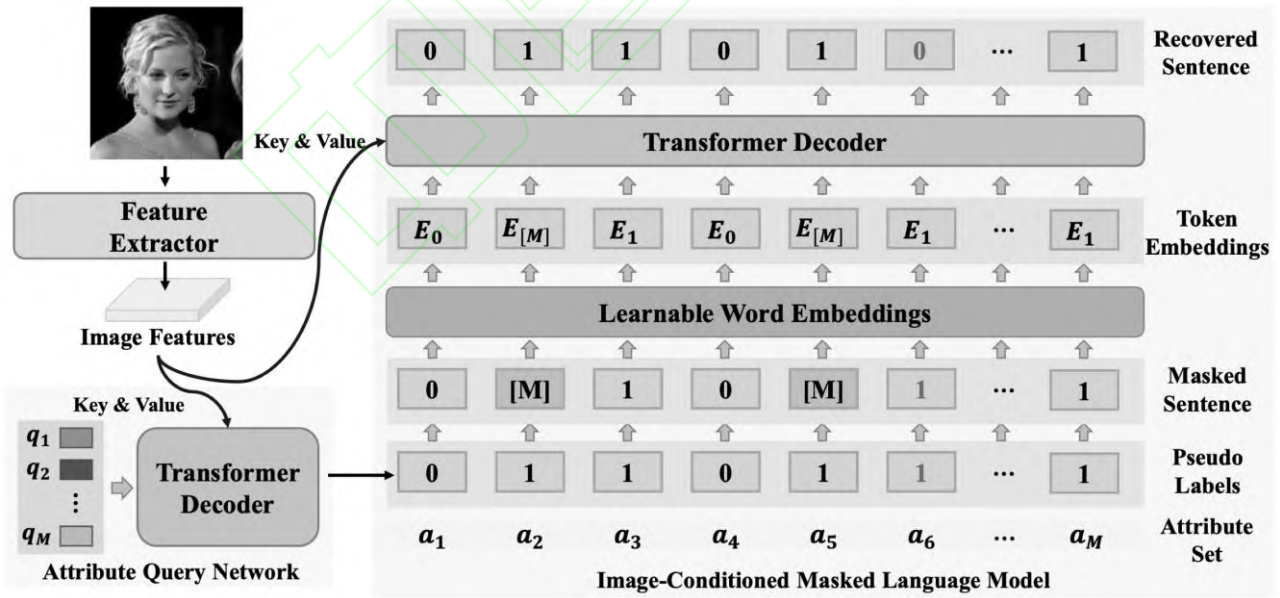

图 22 Label2Label 框架示意图[31]

OAGCN(orientation-aware pedestrian attribute recognition based on GCN). 通常，基于单帧图像识别行人属性的方法忽视了行人朝向对属性的可见性以及相关性的影响，导致属性空间注意力的同质化，即无论行人处于何种朝向，属性的相关区域定位结果都是类似的，这个现象限制了属性识别的性能.

Lu 等[32]基于 GCN 提出一种 OAGCN 方法，增强了基于行人方向的属性空间注意区域，建立了方向引导的属性关系学习.

如图 23 所示, OAGCN 方法由朝向感知空间注意力(orientation-aware spatial attention, OSA)模块

和朝向引导属性关系学习(orientation-guided attribute-relation learning, OAL)模块组成．其中，OSA 模块用于朝向感知特征提取，通过为每个属性生成方向感知的空间注意力构造特征激活图的伪序列，并通过 LSTM 构建校正器，使 OSA 模块能够更好地集中于属性区域并增强视觉属性的学习表示；OAL 模块在 GCN 的基础上提取和判别每个朝向的内在属性关系，通过图卷积将属性的特征作为图节点进行源传播和目标传播，针对不同朝向构建不同的属性关系．此外，该方法首次将 swin Transformer 作为骨干网络应用于行人属性识别领域，有效地提升了属性识别结果．

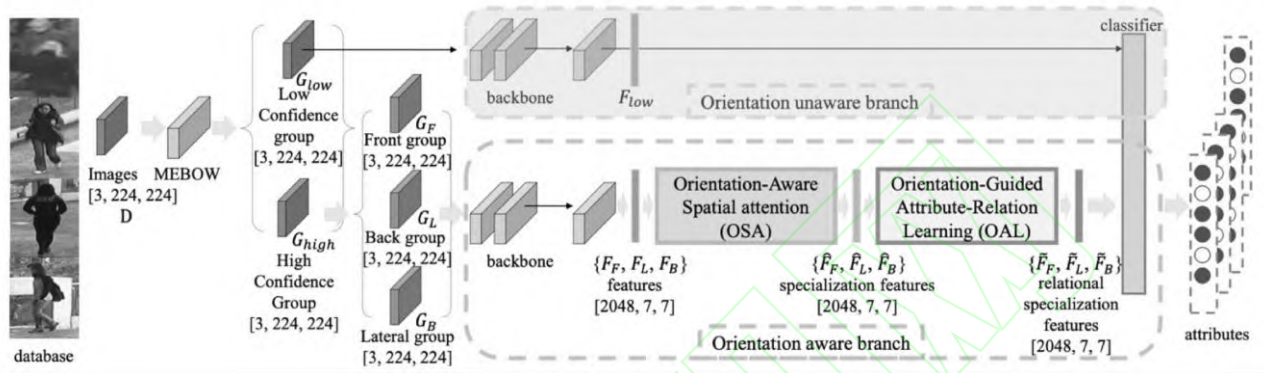

图 23　OAGCN 方法的结构[32]

OAGCN 方法有助于理解行人状态，并为其他行人任务提供帮助，但该方法没有研究更多行人状态对 PAR 的可能影响．

通常，属性之间存在一定的关联性和相互依赖关系，通过挖掘属性之间的关系，可以利用上下文信息增强属性预测的准确性．除了属性之间可能存在的关联性与依赖性，某些属性的存在也可为其他属性的预测提供线索，基于此，利用属性关系进行合理推测有助于判断图片中不可视区域以及画面模糊区域的属性；研究者们通过使用 GCN 生成属性关系图等方式构建属性之间的关系模型，提升行人属性识别的准确率．然而，利用属性关系的方法在识别过程中存在一定的负面影响，例如，“长发”属性可以指根据行人头部、肩部位置的特征信息进行识别判断，而大量女性和长发同时出现的样本使得此类方法倾向于将这 2 个属性捆绑识别，形成“刻板关系”，即当识别某一行人为女性时，无论该行人的发型实际是否为长发，长发属性的输出概率都会提升．此外，数据集的场景限定性也易使识别模型对某一数据集中特定属性的共存关系过拟合，不利于泛化性能的提升．因此，如何恰当利用属性之间的关系，避免刻板关系对属性识别结果造成负优化，是构建属性关系模型中需要关注的问题．

2.2　基于序列图像的行人属性识别方法

不同于基于单帧图像的行人属性识别方法，基于序列图像的方法通过融合同一行人的多帧图像，获取该行人在不同朝向、不同视角下的属性信息，弥补单帧图像下视角和行人姿态固定的局限性，提升单帧图像中不可见区域的属性识别准确率，得到该行人更为完整的属性信息．

2.2.1　融合时间注意力机制

时间注意力策略模型．Chen 等[33]首次提出基于视频的行人属性识别方式，在传统神经网络的基础上，提出一种基于时间注意力策略的多任务模型．由于此前使用的数据集大多是基于独立图片构建的，不能满足基于视频的行人属性识别方法对输入数据的要求，Chen 等对基于视频构建的 MARS 数据集中的样本进行属性标注，得到 MARS-attribute 数据集，并将这些属性分为 2 类：一类是行为相关属性，包括动作和姿态；另一类是身份相关属性，包括服饰颜色、性别等．MARS-attribute 数据集的属性如表 2 所示．

表 2　MARS-attribute 数据集属性分类

属性类别		属性名称	识别结果
行为相关	多元属性	动作	走，站，跑，骑行，变化
		姿态	正面，侧前，侧，侧后，背面，变化
身份相关		年龄	儿童，青少年，成年人，老年人

	上衣颜色	黑，白，粉，紫，黄，灰，蓝，绿，棕，混合
	下装颜色	黑，白，红，紫，黄，灰，蓝，绿，混合
二元属性	性别	男，女
	头发长度	长，短
	上衣长度	长，短
	下装长度	长，短
	下装类型	裤，裙
	帽子	是，否
	背包	是，否
	单肩包	是，否
	手提包	是，否

视频处理方法主要面临的挑战是如何有效融合并利用行人的时空信息. 如图 24 所示, Chen 等[33]提出的模型分为 2 个与属性分类相对应的通道: 通道 1 处理与身份相关的属性; 通道 2 处理与行为相关的属性. 该模型以同一行人的多帧图片作为输入, 将 Resnet-50 作为骨干网络提取每帧图片的空间特征; 之后, 这些空间特征将输入 2 个通道进行属性识别处理, 提出时间注意力模块来确定每帧的重要程度, 从而对同一行人在不同时间点上的图像给予不同的关注权重; 最后由空间特征向量和时间特征向量共同作用, 得到某一属性的特征向量并将其用于属性分类, 得到最终识别结果.

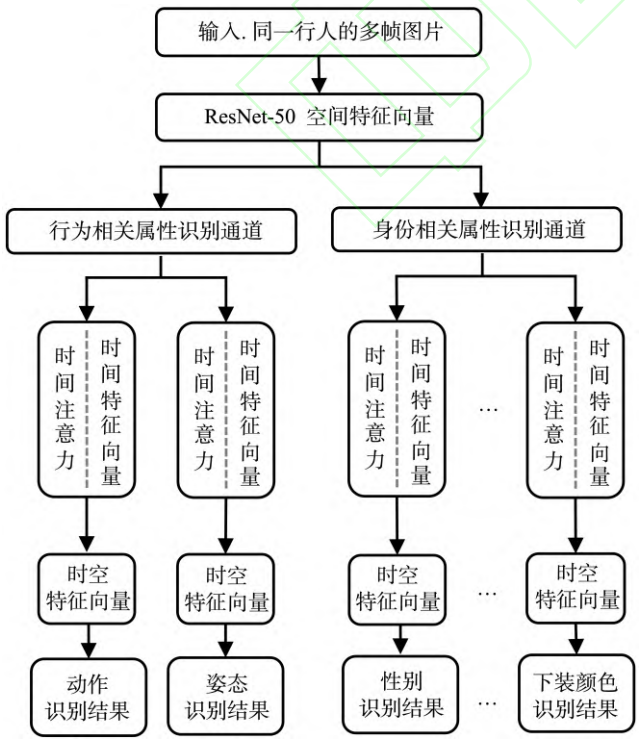

图 24 时间注意力策略模型示意图

由于不同属性在识别过程中依赖的帧可能不同, 该模型为每个属性都配备了独立的时间注意模块, 削减了每帧图片共享同一空间特征向量这个步骤带来的负面影响.

STAM(sigmoid-based temporal attention module). 在视频中, 由于行人处于运动状态, 该时刻处于遮挡的人体部分可能会在另一时刻处于可见状态. 为了利用视频中丰富的信息, Lee 等[34]延续了基于视频利用同一行人多帧图片联合识别属性[33]的思路, 提出一种基于群稀疏性的时间注意力模块处理遮挡问题, 旨在引导属性识别模型关注无遮挡帧.

Chen 等[33]提出的时间注意力模块是基于 softmax 函数构建的概率模型, 该模型中的 ReLU-softmax 单元得到的注意力权重并不能准确地反映稀疏性约束. 为了解决这一问题, Lee 等[34]运用 sigmoid 函数构建新的注意力模块 STAM, 以更好地提取时间特征向量, 其结构图如图 25 所示.

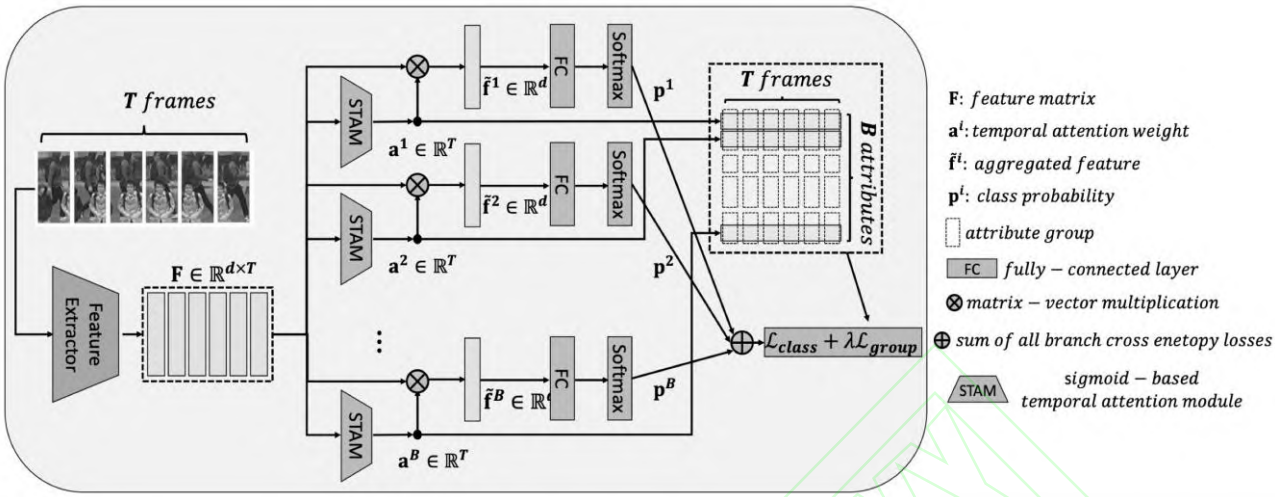

图 25　STAM 的结构[34]

此外，Chen 等[33]提出的基于时间注意力策略的多任务模型中将属性按照行为相关和身份相关分为 2 大类，在 2 类属性的识别通道内独立处理各属性，忽略了属性之间客观存在的联系．为此，Lee 等[34]做出了优化，按照属性与空间之间的关系将其分为 5 组．由于各组内的属性之间存在一定的联系，如“鞋子类型”和“鞋子颜色”均与行人足部相关，因此同组的属性会获得相近的注意力权重．STAM 中属性分组如表 3 所示．

表 3　STAM 属性分组

属性组别	相关区域	属性名称
1	整体	动作，姿态
2	头部	帽子，性别
3	上身	背包，上衣颜色，单肩包，手提包
4	下身	上衣长度，下装颜色
5	足部	鞋子类型，鞋子颜色

STAM 中同样使用 ResNet 提取空间特征；之后将其输入 STAM，根据属性赋予不同帧相应的注意力权重，引导模型关注无遮挡的帧来识别属性，对得到的特征向量进行增强；最后利用多分支分类器得到属性的判断结果．

2.2.2　改进注意力机制及损失函数

MTA-Net(multiple time attention network)．属性和图片之间的复杂关系以及属性的不平衡分布是行人属性识别任务中的 2 个重要问题，CNN-RNN 和注意力机制已被广泛用于挖掘属性与图片之间的关系，而为了弥补属性分布不平衡给模型结果带来的偏差，研究者们也提出多种损失函数进行调整．Ji 等[35]引入新的注意力机制和损失函数，并将两者融合，构建了 MTA-Net，其结构如图 26 所示．

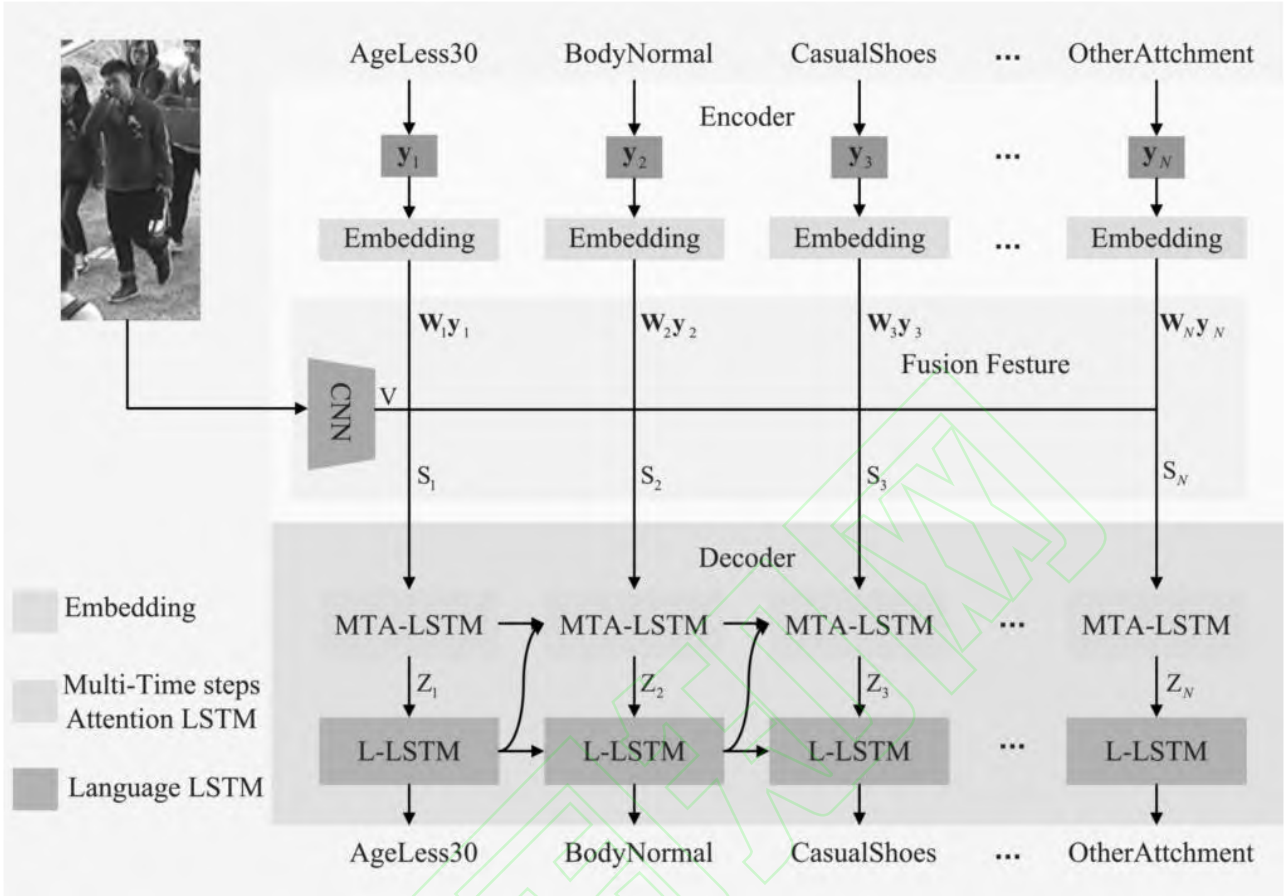

图 26 MTA-Net 的结构[35]

常见的 CNN-RNN 模型和注意力机制通常只关注已出现的的前序时间和当前所在的时间，忽略了后续时间序列中所包含的信息. 为此，Ji 等[35]在 LSTM 的基础上提出 MTA 机制，考虑之前、当下和之后 3 段时间状态，能够更全面地探索属性和图像之间的关系；此外，还引入焦平衡损失函数(focal balance loss, FBL)来解决属性分布不平衡问题，计算公式为

$$\mathrm{FBL}=-\left[\omega_i+(1-p_i)+\omega_i(1-p_i)\right]\sum_i \lg(p_i).$$

其中，$\omega_i=\exp(-r)$，r 表示训练集中第 i 个属性的正样本所占比例；p_i 表示第 i 个属性输出的可能性.

因此，与单帧图像相比，序列图像包含更丰富的时序及上下文信息，这类行人属性识别方法致力于利用同一行人的多帧图片挖掘更多属性信息，可以通过不同时间、不同视角下拍摄的图片弥补单帧图片中不可视区域的信息缺失. 时空注意力机制可以在不同时间和空间尺度上进行加权，以更好地捕捉行人属性的时空特征，从而获取同一行人更为准确的属性识别结果.

目前，基于序列图像的行人属性识别方法仍处于发展初期，且适用于该类方法的数据集较少，因此该方向的行人属性识别方法有很大的发展空间.

3 数据集及现有方法的评估

3.1 常用数据集

不同于人体属性识别数据集 HAT[36]等，行人属性识别数据集中图片的主体部分仅为某一位行人，且图像边界为该行人的边界框，包含的背景内容较少；而人体属性识别所用的数据集对图像的采集场景并没有特定要求，其画面中可能包含一位或多位人物，且不要求其处于行走的运动状态.

目前，行人属性识别领域常用的数据集概览如表 4 所示.

表 4 行人属性识别领域常用数据集概览

数据集	样本数量	行人数量	二元属性数量	多元属性数量	拍摄场景	相机状态	特点

PETA	19000	8705	61	4	室内+室外		汇集10个较小规模的数据集，样本类型丰富，基于行人身份进行标注
RAP-v1	41585		69	3	室内	静止	样本、属性标注丰富，均采集自室内，光线条件和背景较为单一，基于图片进行标注
PA-100K	100000		26	0	室外		目前为止规模最大，按照行人身份划分训练集和测试集，比例为8∶1∶1，基于图片进行标注
PARSE-27k	27000		8	2	室外	移动	图片截取自8段城市环境下的视频，引入N/A标注表示难以判断的属性按照视频时序划分训练集和测试集，比例为2∶1∶1
Market-1501-attribute	32668	1501	27	0	室外	静止	样本采集于夏季，服装类型集中在夏装，基于行人身份进行标注
MARS	16360	1261	9	5	室外	静止	Market-1501的扩展版本

3.1.1　PETA数据集

PETA数据集[6]发布于2014年，由10个较小规模的数据集集合而得，图像分辨率为17×39~169×365像素不等，共包含19000份样本，涵盖8705位行人，包含室内和室外场景，样本共被标注了61个二元属性和4个多元属性. 与早期的数据集相比，该数据集的样本采集视角丰富，涵盖多种行人朝向，光照条件多变，且行人所处的背景环境更为复杂多样；但是，在对某一行人进行属性标注时，PETA数据集采用的方法是随机抽取该行人的一帧图像进行标注并将其应用于该行人的所有图像，存在属性标注不准确、不完善的情况，并且采取随机划分的方式得到训练集和测试集存在数据泄露问题.

3.1.2　RAP数据集

RAP数据集[37]第1版(RAP-v1)发布于2016年，包含41585份收集自室内商场监控的样本；这些样本共被标注了72个属性，其中包括3个多元属性和69个二元属性；标注属性时，不仅考虑行人自身，还考虑了环境因素，对行人的采样视角和遮挡类型进行标注，进一步丰富了属性标签. 与PETA数据集的划分方式相同，RAP数据集也存在随机划分训练集和测试集致使数据泄露的问题.

3.1.3　PA-100K数据集

Liu等[20]建立了PA-100K(pedestrian attribute-100K)数据集，其中包含100000张采集自600个真实室外场景的行人图片，是目前规模最大的行人属性识别数据集；构建数据集时，剔除过于模糊及分辨率过低的图片，提高了样本图片质量，对样本标注26个常用属性，按照行人身份将样本以8∶1∶1的比例划分为训练集、验证集和测试集，避免了训练集和测试集中行人身份重合的现象.

3.1.4　PARSE-27k数据集

Sudowe等[38]发布了PARSE(pedestrian attribute recognition on sequences)-27k数据集，其在8段城市环境下的视频中截取了27000张行人图片，按视频时序以2∶1∶1的数据比例构建训练集、验证集和测试集，减少了高度相似的图片出现在不同集合中的可能性. 不同于之前的标注方式，该数据集在标注时引入N/A标注，用于表示难以判断的属性. 然而，该数据集的属性类别较少，仅涵盖8个二元属性和2个多元属性.

3.1.5　Market-1501-attribute数据集

2015年，Zheng等[39]发布了Market-1501数据集，其中的行人图片采集自清华大学校园超市前的6个监控摄像头，包含1501位行人的32668张图片. 该数据集的样本采集于夏季，服装类型集中在夏装，具有一定的场景局限性. 为了利用行人属性识别来提升行人重识别结果，Lin等[40]基于行人身份对PARSE-27k数据集中的样本进行属性标注，构建了Market-1501-attribute数据集，每张图片均被标注了27个属性.

3.1.6　MARS-attribute数据集

MARS数据集[41]是Market-1501数据集的拓展版本，通过6个摄像头采集了1261位行人的图片. Chen等[33]以Lin等[40]所采用的属性列表为参考对MARS数据集进行属性标注建立MARS-attribute数据集，在外观特征等属性的基础上，增加了动作和姿态2项描述行人状态的属性，将属性分为身份相关和行为相关2大类，共包含14个属性，其中，9

个为二元属性, 5 个为多元属性, 如表 2 所示.

上述数据集中, RAP-v1, PA-100K 和 PARSE-27k 是面向行人属性识别任务直接构建的数据集; PETA, Market-1501-attribute 和 MARS-attribute 数据集则是通过对行人再识别数据集标注行人属性而来. 其中, PETA 数据集中的部分子集(如 MIT, SARC3D)中的行人图片与面向行人属性识别任务直接构建的数据集类似, 如图 27a, 同一行人在不同图像中所处的状态相对独立; 而其部分子集(如 CAVIAR4REID, TownCentre)中的行人图像则与 Market-1501-attribute, MARS 数据集更为类似, 如图 27b, 同一行人的不同图像可构成较为连续的序列.

a. 相对独立的行人图像

b. 较为连续的行人序列图像

图 27 不同类型行人图像样本示例

因此, 基于单帧图像的行人属性识别的方法大多使用 PETA, RAP 和 PA-100K 数据集进行实验结果的评估, 而基于序列图像的方法则通常使用 Market-1501-attribute 和 MARS-attribute 数据集.

当前, 已有的数据集对于行人属性的标注方式可分为基于行人进行标注和基于图片进行标注 2 类. 基于行人的标注方式指同一行人的不同图片共用同一套属性标签, 这种标注方式更适用于基于序列图像识别行人属性的方法, 便于对同一行人生成整体性属性描述, 但对基于单帧图像的识别方法易带来数据噪声的负面影响; 基于图片的标注方式则根据不同样本的具体情况进行标注, 同一行人的不同图片的标注列表可能有所不同, 该种标注方式更适合基于单帧图像的方法.

总的来说, PETA 和 RAP 数据集拥有丰富的属性标签, PA-100K 数据集具有较大的数据规模, PARSE-27k 数据集引入 N/A 的属性标注方式, Market-1501-attribute 和 MARS-attribute 数据集为基于序列图像的行人属性识别方法提供了更为合适的数据. 然而, 当前数据集仍普遍存在具有场景限定性和数据分布不平衡的问题.

3.2 现有方法在常用数据集上的评估

通常, 采用基于属性评估的平均准确率(mean accuracy, MA), 基于样本评估的准确率 A_{exam}、精确率 P_{exam} 和召回率 R_{exam}, 以及 F_1 值这 5 项指标, 对行人属性识别实验结果进行评估. 计算公式为

$$\text{MA}=\frac{1}{2C}\sum_{i=1}^{C}\left(\frac{\text{TP}_i}{P_i}+\frac{\text{TN}_i}{N_i}\right).$$

其中, C 表示样本总数; 对于第 i 个属性来说, P_i 和 N_i 分别表示正向样本和负向样本的数量; TP_i 和 TN_i 分别是识别正确的正向样本(true positive, TP)和识别正确的负向样本(true negative, TN)的数量. 该方法对每个属性的识别准确率都进行单独计算, 再通过求取平均值, 避免了部分属性样本数据较少导致的准确率虚高问题.

$$A_{\text{exam}}=\frac{1}{N}\sum_{i=1}^{N}\frac{\left|Y_i\cap f\left(x_i\right)\right|}{\left|Y_i\cup f\left(x_i\right)\right|},$$

$$P_{\text{exam}}=\frac{1}{N}\sum_{i=1}^{N}\frac{\left|Y_i\cap f\left(x_i\right)\right|}{\left|f\left(x_i\right)\right|},$$

$$R_{\text{exam}}=\frac{1}{N}\sum_{i=1}^{N}\frac{\left|Y_i\cap f\left(x_i\right)\right|}{\left|Y_i\right|},$$

$$F_1=\frac{2\times P_{\text{exam}}\times R_{\text{exam}}}{P_{\text{exam}}+R_{\text{exam}}}.$$

其中, N 表示图片总数; $|\cdot|$ 表示集合 $\cdot$ 的基数; 对于第 i 张图片, Y_i 表示标注为正的属性, $f(x_i)$ 表示被判断为正的属性.

表 5 所示为 16 种方法在 PETA, RAP 和 PA-100K 数据集下的实验结果对比; 表 6 所示为 4 种使用 ResNet-50 骨干网络的方法在 MARS-attribute 和 Market-1501-attribute 数据集下的实验结果对比. 从表 1, 表 5~表 6 可以看出, ResNet 是目前主要使用的骨干网络, 注意力机制是广泛使用的处理技巧; 对于基于单帧图片的识别方法, 引入空间注意力以及挖掘属性之间的联系能够显著地提升行人属性识别的准确率; 对于基于视频多帧图片的识别方法, 融合时间注意力是提高行人属性识别表现的有效手段.

表 5 3 个数据集下 16 种方法实验结果对比 %

方法	骨干网络	PETA					RAP				
		MA	A_{exam}	P_{exam}	R_{exam}	F_1	MA	A_{exam}	P_{exam}	R_{exam}	F_1
P-Net[17]	GoogLeNet	95.18	94.27	95.83	96.81	96.32	84.63	60.85	74.72	74.49	74.60
DTM[18]	ResNet-50	85.24	79.26	86.81	86.67	86.48	81.25	68.60	79.91	81.17	80.53
DTM+AWK[18]	ResNet-50	85.79	78.63	85.65	87.17	86.11	82.04	67.42	75.87	84.16	79.80
SSC_{soft}[19]	ResNet-50	86.52	78.95	86.02	87.12	86.99	82.77	68.37	75.05	87.49	80.43
SSC_{hard}[19]	ResNet-50	85.92	78.53	86.31	86.23	85.96	82.14	68.16	77.87	82.88	79.87
SSC_{fix}[19]	ResNet-50	86.07	79.23	84.58	89.26	86.54	82.83	68.16	74.74	87.54	80.27
Semantic parsing-PAR[23]	EfficientNet	87.69	81.20	87.59	89.20	88.32	82.37	69.93	80.46	87.23	82.33
HR-Net[25] (对部分属性)	ResNet-50	91.46	79.85	84.91	90.43	87.59	82.18	51.32	67.74	65.94	66.84
	GoogLeNet	94.42	94.68	96.47	96.11	96.29	81.10	45.70	51.48	78.56	62.20
DAFL[26]	ResNet-50	87.07				86.40	83.72				80.29
JRL[27]	AlexNet	85.67		86.03	85.34	85.42	77.81		78.11	78.98	78.58
JLAC[28]	ResNet-50	86.96	80.38	87.81	87.09	87.45	83.69	69.15	79.31	82.40	80.82
CGCN[30]	ResNet-101	84.70	54.50	60.03	83.68	70.49					
Label2Label[31]	ResNet-50										
OAGCN[32]	Swin transformer	90.11	81.25	86.84	88.99	87.90	87.93	69.97	77.03	86.82	81.63
MTA-Net[35]	ResNet-152	84.62	78.80	85.67	86.42	86.04	77.62	67.17	79.72	78.44	79.07

方法	骨干网络	PA-100K				
		MA	A_{exam}	P_{exam}	R_{exam}	F_1
P-Net[17]	GoogLeNet					
DTM[18]	ResNet-50	80.70	78.82	87.37	87.02	87.20
DTM+AWK[18]	ResNet-50	81.63	77.57	84.27	89.02	85.68
SSC_{soft}[19]	ResNet-50	81.87	78.89	85.98	89.10	86.87
SSC_{hard}[19]	ResNet-50	81.02	78.42	86.39	87.55	86.55
SSC_{fix}[19]	ResNet-50	81.70	78.85	85.80	88.92	86.89
Semantic parsing-PAR[23]	EfficientNet	81.45	79.13	86.24	89.46	87.94
HR-Net[25] (对部分属性)	ResNet-50					
	GoogLeNet					
DAFL[26]	ResNet-50	83.54				88.09
JRL[27]	AlexNet					-
JLAC[28]	ResNet-50	82.31	79.47	87.45	87.77	87.61
CGCN[30]	ResNet-101					
Label2Label[31]	ResNet-50	82.24	79.23	86.39	88.57	87.08
OAGCN[32]	Swin transformer	84.07	80.54	83.18	89.00	85.99
MTA-Net[35]	ResNet-152					

表 6　使用 ResNet-50 在 2 个数据集下 4 种方法实验结果对比　　%

数据集	方法	MA	A_{exam}	P_{exam}	R_{exam}	F_1
Market-1501-attribute	HFE[29]		78.01	87.41	85.65	86.52
MARS-attribute	时间注意力策略模型[33]	89.31				73.24
	STAM(对遮挡样本)[34]	71.94				61.88

	STAM(对所有样本)[34]	86.75	70.42

然而，在某个数据集下取得较好结果的方法在另一个数据集下不一定能够达到同样理想的效果. 不仅如此，Jia 等[42]按照零样本的设定构建了 PETA_{ZS} 和 RAP_{ZS} 数据集，以满足 $Z_{\text{train}} \bigcap Z_{\text{test}} = \varnothing$ 的条件，其中，Z_{train} 和 Z_{test} 分别表示训练集和测试集中行人身份，即训练集和测试集中的行人身份互不重叠，而原始的 PETA 和 RAP 数据集中，训练集和测试集随机划分，同一行人的不同图片可能会同时出现在训练集和测试集中；他们还将 JLAC[28]等方法分别在 PETA 和 RAP 数据集、PETA_{ZS} 和 RAP_{ZS} 数据集上进行实验，结果如表 7 所示. 可以看出，在 PETA_{ZS} 和 RAP_{ZS} 数据集上，行人属性识别方法的各项性能指标均有下降. 因此，行人属性识别方法的泛化能力仍需进一步提升.

表 7 JLAC 方法[28]在 4 个数据集下实验结果[42]对比 %

数据集	MA	A_{exam}	P_{exam}	R_{exam}	F_1
PETA	86.02	79.51	86.62	87.19	86.66
PETA_{ZS}	73.60	58.66	71.70	72.41	72.05
RAP	79.23	64.42	75.69	79.18	77.40
RAP_{ZS}	76.38	62.58	73.14	79.20	76.05

4 未来研究方向

4.1 状态引导行人属性识别

行人状态包括朝向、动作、姿态、遮挡等多种信息，当行人状态发生变化时，其属性的空间关系及属性之间的相关性会随之发生变化. 因此，在行人属性识别中增加状态引导的思想，有助于提升识别效果.

同一目标行人在不同朝向、不同动作、不同姿态下，可直接观测到的属性有所不同，行人肢体的空间位置也会发生改变，从而使属性相关区域在行人图像中所处位置随之改变. 通过划分行人朝向、使用姿态估计等手段建模行人动作与姿态，可以作为行人属性识别的引导，提升属性相关区域定位的准确度，优化属性之间共现关系与互斥关系的构建.

在此基础上，针对遮挡问题，可发展基于状态引导融合多帧行人图像的识别方法，在目标行人状态发生变化时，使不同状态下的帧的识别结果形成互相补充关系并进行交叉验证；对在多数帧中出现的属性做出准确判断，并对在少数帧中出现的属性做出可信判断，从而对遮挡区域的属性作出合理推断.

4.2 构建立体属性

构建立体属性是基于序列图像识别行人属性的扩展思路. 基于单帧图片识别行人属性的方法孤立了同一行人不同朝向下的属性判断结果，混淆了属性的“不可见”和“不存在”状态，易造成行人属性识别结果不够全面. 因此，构建立体属性能够从多视角更为完善地描述行人的属性.

例如，对于行人的“背包”属性，若以 1, 0 分别表示该属性的存在与否，当一位背包的行人正对摄像头时，该属性易被判断为 0，而当该行人侧对、背对摄像头时，该属性易被判断为 1；针对该行人进行背包这一属性的预测时，这 2 种识别结果则会互相干扰，若该属性在这位行人的大多数图像中均处于被遮挡状态或该行人在大多数图片中朝向为正，则易使得背包的最终判断结果为 0，从而导致该属性信息的丢失. 即当该行人的单帧图像被单独处理时，其属性判断结果是基于某一视角得到的不完整的、平面化的结果.

立体属性的概念是基于多帧图片序列提出的拓展，对于同一行人，提取其多帧图片进行属性识别，根据不同属性与各朝向的关联程度，对不同朝向下的行人图像赋予不同的关注程度，最终得到综合判断结果，实现不同朝向、不同视角下的图片的空间互补. 将平面化的属性识别结果发展为更为完整的立体化属性，解决视角变化、遮挡等问题带来的困难.

如图 28 所示，对于单一朝向下的行人图像，仅能得到与之对应的正向视角下的属性，其信息局限于一个二维平面内；综合考虑多朝向下的行人图片，有助于在三维空间的视角下获取更为全面的行人属性信息，从而对这一行人生成更为准确的结构化描述.

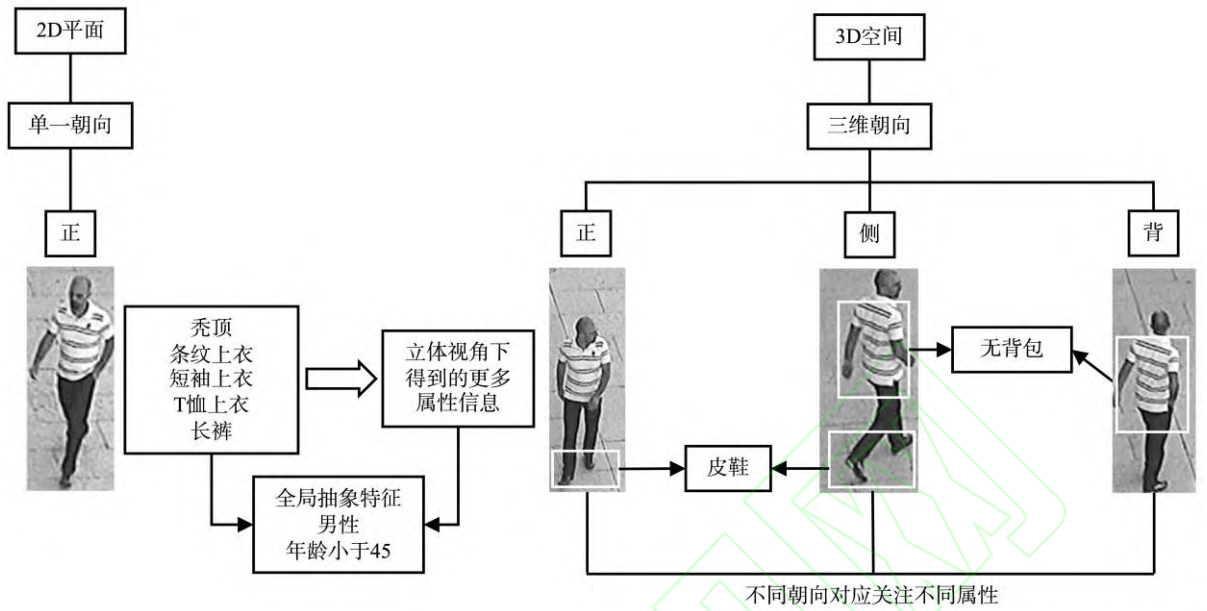

图 28　立体属性构建示例

4.3　多任务融合

在计算机视觉与安防监控的应用领域，行人属性识别和行人重识别是 2 项相似且具有一定联系的 2 项任务.

行人重识别被广泛认为是图像检索的一个子问题，旨在跨设备情境下判断图像或视频序列中是否存在某一个给定的行人，从而弥补固定摄像头的视觉局限. 行人属性识别的目标是判断某一给定行人的特征，包括性别、穿着风格、长发、皮鞋、背包等属性，从而生成该行人的结构化描述.

从任务目标看，行人重识别是确定行人身份，而行人属性识别是确定描述该行人特征的属性信息；然而，这 2 项目标不同的任务在实际运用中可以互相辅助，带有行人身份标注的数据集 Market-1501-attribute 和 MARS-attribute 也为该思路提供了数据支持. 如图 29 所示，在行人重识别过程中，利用行人属性识别得到多帧图片下行人具有相同属性，则这些图片中的行人属于同一身份的概念更大；在行人属性识别的过程中，多帧图片中的行人被判断为同一身份时，这些图片中的行人客观上具有相同的属性信息，可辅助与遮挡区域的属性推测.

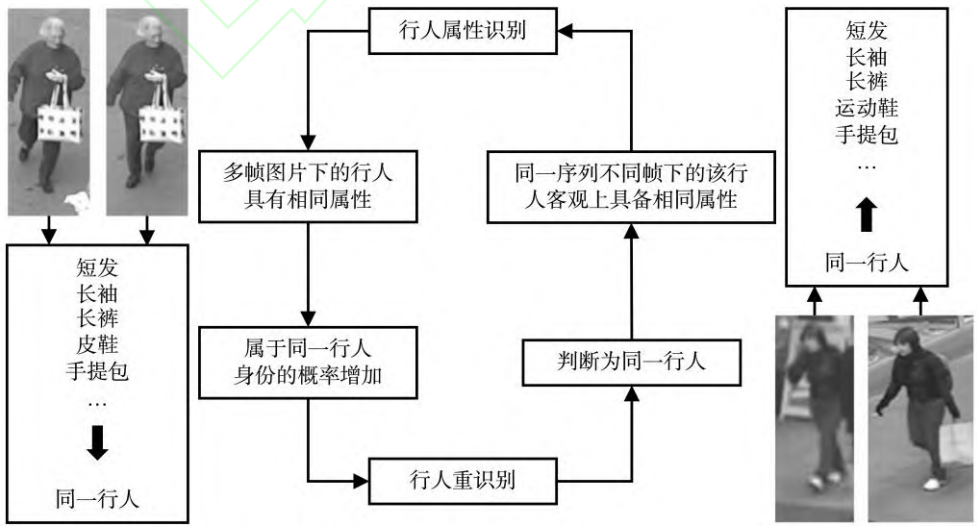

图 29　行人属性识别与行人重识别融合

4.4　构建新数据集

已有的用于训练行人属性识别算法模型的数据集中还存在不足之处，会对最终识别效果的准确性产生影响：训练集和测试集之间的数据泄露问题有可能导致实验结果虚高，而数据偏置现象则易造成泛化性能较弱等问题.

同一数据集中的行人图像大多采集自同一场景或相似场景，存在场景限定性，而该现象也易导致数据集中的属性分布存在一定的偏置现象，出现属性数据分布不均衡的问题．不仅如此，在 PETA，RAP 等数据集中，训练集和测试集的划分标准较为随机，导致同一行人的图像样本可能会在训练集和测试集中同时出现，对属性识别的结果造成影响．此外，以 0, 1 标注二元属性的标注方式也混淆了属性“存在但出于不可见区域”和“不存在”的 2 种现象．

针对上述问题，PARSE-27k 数据集在标注属性时引入 N/A 标注，用于表示难以判断的属性; Jia 等[42]按照零样本的设定，构建了 $PETA_{ZS}$ 和 RAP_{ZS} 数据集，使得训练集和测试集中的行人身份互不重叠; 为了探究行人属性识别方法的泛化能力，Specker 等[43]融合了 PA-100K，PETA，RAPv2 和 Marke-t1501 这 4 个数据集，对其中的行人图片进行统一标注，构建了 UPAR 数据集，共有 224 737 张标有 40 个二元属性的行人图片，这些属性包含年龄、性别等全局属性，以及眼镜、帽子、上下装颜色等部件级局部属性．该数据集融合了多个较大规模的现有数据集，使其样本量显著增加，以服饰、人种、场景的多样性削弱了单一数据集的局限性，并在一定程度上缓解了属性分布不均衡的问题; 其属性标注主要集中于年龄、性别、发型、配饰以及服装颜色，忽略了对行人状态的标注，即行人朝向、动作等信息．

构建合理的新数据集能够有效地提升行人属性识别的性能，而当前使用的数据集仍存在一些的不合理之处，基于此，理想的行人属性识别数据集应满足以下条件:

(1) 场景广泛．数据集中的图像应采集自时空条件不同的各类环境，丰富样本类型，以缓和场景限定性问题;

(2) 数据均衡．如文献[6]中的定义，对于二元属性, 2 个分类中的样本数量之比不应超过 20∶1, 从而减少对高频属性过拟合现象的发生;

(3) 标注多样．属性标注类别中可以引入“−1”的标注，表示不可视区域的属性，将其与能够在可视区域明确识别为“不存在”的属性进行区分; 此外，在考虑行人年龄、性别、服饰特征的基础上，融入行人朝向等状态信息的标签，引导属性识别．

5 结 语

本文面向行人属性识别领域，围绕最新技术发展下的研究热点与关键挑战，从基于单帧图像和基于序列图像 2 个方向，根据方法机制对行人属性识别方法进行分类以及归纳总结，分析对比了基于图片和基于视频 2 类常用数据集以及现有方法的实验评估结果; 同时，面向探索状态引导行人属性识别、构建立体属性、挖掘属性识别融合上下游多任务的联合学习范式、构建新数据集等研究热点，展望了该领域的未来发展，以期对该领域的研究人员有所裨益．

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] //Proceedings of the IEEE International Conference on Computer Vision Workshops. Los Alamitos: IEEE Computer Society Press, 2013: 331-338, doi: 10.1109/ICCVW.2013.51

[2] Wang X, Zheng S F, Yang R, *et al*. Pedestrian attribute recognition: a survey[J]. Pattern Recognition, 2022, 121: Article No.108220

[3] Yaghoubi E, Khezeli F, Borza D, *et al*. Human attribute recognition— a comprehensive survey[J]. Applied Sciences, 2020, 10(16): Article No.5608

[4] Jia Jian, Chen Xiaotang, Huang Kaiqi. Pedestrian attribute recognition in surveillance scenes: a survey[J]. Chinese Journal of Computers, 2022, 45(8): 1765-1793(in Chinese)
(贾健, 陈晓棠, 黄凯奇. 监控场景中的行人属性识别研究综述[J]. 计算机学报, 2022, 45(8): 1765-1793)

[5] Zhang H D, Gao Y Y, Hu H M, *et al*. Single image haze removal based on global-local optimization for depth map[C] //Proceedings of the 18th Pacific-Rim Conference on Multimedia on Advances in Multimedia Information Processing. Heidelberg: Springer, 2018: 117-127

[6] 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. New York: ACM Press, 2014: 789-792, doi: 10.1145/2647868.2654966

[7] Tang C, Sheyng L, Zhang Z, *et al*. Improving pedestrian attribute recognition with weakly-supervised multi-scale attribute-specific localization[C] //Proceedings of the IEEE/CVF International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2019: 4996-5005

[8] Li D W, Chen X T, Huang K Q. Multi-attribute learning for pedestrian attribute recognition in surveillance scenarios[C] //Proceedings of the 3rd IAPR Asian Conference on Pattern Recognition (ACPR). Los Alamitos: IEEE Computer Society Press, 2015: 111-115, doi: 10.1109/ACPR.2015.7486476

[9] Guo H, Zheng K, Fan X C, *et al*. Visual attention consistency under image transforms for multi-label image classification[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2019: 729-739

[10] Li Q Z, Zhao X, He R, *et al*. Visual-semantic graph reasoning for pedestrian attribute recognition[C] //Proceedings of the 33rd AAAI Conference on Artificial intelligence. Palo Alto: AAAI Press, 2019: 8634-8641

[11] Wu J J, Liu H, Jiang J G, *et al*. Person attribute recognition by

sequence contextual relation learning[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2020, 30(10): 3398-3412
[12] Bourdev L, Maji S, Malik J. Describing people: a poselet-based approach to attribute classification[C] //Proceedings of the International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2011: 1543-1550, doi: 10.1109/ICCV.2011.6126413
[13] Zhang N, Paluri M, Ranzato M, *et al*. PANDA: pose aligned networks for deep attribute modeling[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2014: 1637-1644
[14] Zhu J Q, Liao S C, Yi D, *et al*. Multi-label CNN based pedestrian attribute learning for soft biometrics[C] //Proceedings of the International Conference on Biometrics. Los Alamitos: IEEE Computer Society Press, 2015: 535-540
[15] Zhou B L, Khosla A, Lapedriza A, *et al*. Learning deep features for discriminative localization[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2016: 2921-2929
[16] Zheng L, Huang Y J, Lu H C, *et al*. Pose invariant embedding for deep person re-identification[J]. IEEE Transactions on Image Processing, 2019, 28(9): 4500-4509
[17] An H R, Fan H N, Deng K W, *et al*. Part-guided network for pedestrian attribute recognition[C] //Proceedings of the IEEE Visual Communications and Image Processing (VCIP). Los Alamitos: IEEE Computer Society Press, 2019: 1-4, doi: 10.1109/VCIP47243.2019.8965957
[18] Zhang J J, Ren P Y, Li J M. Deep template matching for pedestrian attribute recognition with the auxiliary supervision of attribute-wise keypoints[OL]. [2023-06-20]. https://arxiv.org/abs/2011.06798
[19] Jia J, Chen X T, Huang K Q. Spatial and semantic consistency regularizations for pedestrian attribute recognition[C] //Proceedings of the IEEE/CVF International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2021: 942-951
[20] Liu X H, Zhao H Y, Tian M Q, *et al*. HydraPlus-Net: attentive deep features for pedestrian analysis[C] //Proceedings of the IEEE International Conference on Computer Vision (ICCV). Los Alamitos: IEEE Computer Society Press, 2017: 350-359, doi: 10.1109/ICCV.2017.46
[21] Sarafianos N, Xu X, Kakadiaris I A. Deep imbalanced attribute classification using visual attention aggregation[C] //Proceedings of the 15th European Conference on Computer Vision. Heidelberg: Springer, 2018: 708-725
[22] Liu P Z, Liu X H, Yan J J, *et al*. Localization guided learning for pedestrian attribute recognition[OL]. [2023-06-20]. https://arxiv.org/abs/1808.09102
[23] Moghaddam M, Charmi M, Hassanpoor H. Jointly human semantic parsing and attribute recognition with feature pyramid structure in EfficientNets[J]. IET Image Processing, 2021, 15(10): 2281-2291
[24] Tan M X, Pang R M, Le Q V. EfficientDet: scalable and efficient object detection[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2020: 10778-10787
[25] An H R, Hu H M, Guo Y F, *et al*. Hierarchical reasoning network for pedestrian attribute recognition[J]. IEEE Transactions on Multimedia, 2021, 23: 268-280, doi: 10.1109/TMM.2020.2975417
[26] Jia J, Gao N Y, He F, *et al*. Learning disentangled attribute representations for robust pedestrian attribute recognition[C] //Proceedings of the 36th AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2022: 1069-1077
[27] Wang J Y, Zhu X T, Gong S G, *et al*. Attribute recognition by joint recurrent learning of context and correlation[C] //Proceedings of the IEEE International Conference on Computer Vision (ICCV). Los Alamitos: IEEE Computer Society Press, 2017: 531-540, doi: 10.1109/ICCV.2017.65
[28] Tan Z C, Yang Y, Wan J, *et al*. Relation-aware pedestrian attribute recognition with graph convolutional networks[C] //Proceedings of the 34th AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2020: 12055-12062
[29] Yang J, Fan J R, Wang Y R, *et al*. Hierarchical feature embedding for attribute recognition[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2020: 13052-13061
[30] 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, doi: 10.1109/TMM.2020.3045286
[31] Li W H, Cao Z X, Feng J J, *et al*. Label2Label: a language modeling framework for multi-attribute learning[C] //Proceedings of the 17th European Conference on Computer Vision. Heidelberg: Springer, 2022: 562-579
[32] Lu W Q, Hu H M, Yu J Z, *et al*. Orientation-aware pedestrian attribute recognition based on graph convolution network[OL]. [2023-06-20]. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=10078344
[33] Chen Z Y, Li A N, Wang Y H. A temporal attentive approach for video-based pedestrian attribute recognition[C] //Proceedings of the Second Chinese Conference on Pattern Recognition and Computer Vision. Heidelberg: Springer, 2019: 209-220
[34] Lee G, Yun K M, Cho J. Robust pedestrian attribute recognition using group sparsity for occlusion videos[OL]. [2023-06-20]. https://arxiv.org/abs/2110.08708
[35] 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
[36] LeCun Y, Bengio Y, Hinton G. Deep learning[J]. Nature, 2015, 521(7553): 436-444
[37] Li D W, Zhang Z, Chen X T, *et al*. A richly annotated dataset for pedestrian attribute recognition[OL]. [2023-06-20]. https://arxiv.org/abs/1603.07054
[38] Sudowe P, Spitzer H, Leibe B. Person attribute recognition with a jointly-trained holistic CNN model[C] //Proceedings of the IEEE International Conference on Computer Vision Workshop (ICCVW). Los Alamitos: IEEE Computer Society Press, 2015: 329-337, doi: 10.1109/ICCVW.2015.51
[39] Zheng L, Shen L Y, Tian L, *et al*. Scalable person re-identification: a benchmark[C] //Proceedings of the IEEE International Conference on Computer Vision (ICCV). Los Alamitos: IEEE Computer Society Press, 2015: 1116-1124, doi: 10.1109/ICCV.2015.133
[40] Lin Y T, Zheng L, Zheng Z D, *et al*. Improving person re-identification by attribute and identity learning[J]. Pattern Recognition, 2019, 95: 151-161
[41] Zheng L, Bie Z, Sun Y F, *et al*. MARS: a video benchmark for large-scale person re-identification[C] //Proceedings of the 14th European Conference on Computer Vision. Heidelberg: Springer, 2016: 868-884, doi: 10.1007/978-3-319-46466-4_52
[42] Jia J, Huang H J, Chen X T, *et al*. Rethinking of pedestrian attribute recognition: a reliable evaluation under zero-shot pedestrian identity setting[OL]. [2023-06-20]. https://arxiv.org/abs/2107.03576
[43] Specker A, Cormier M, Beyerer J. UPAR: unified pedestrian attribute recognition and person retrieval[C] //Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. Los Alamitos: IEEE Computer Society Press, 2023: 981-990