监控场景中的行人属性识别研究综述

贾 健[1),2)] 陈晓棠[1),2)] 黄凯奇[1),2),3)]

[1)](中国科学院大学人工智能学院, 北京 100049)

[2)](中国科学院自动化研究所智能系统与工程研究中心 北京 100190)

[3)](中国科学院脑科学与智能技术卓越创新中心 上海 200031)

摘 要 监控场景中的行人属性识别任务旨在为监控场景中视频摄像头捕捉的行人图片预测其属性类别，由于监控场景环境的复杂以及行人属性的细粒度标签，监控场景中的行人属性识别任务极具挑战，受到业界和学界的广泛关注。本文对监控场景中的行人属性识别研究进展进行梳理，首先给出了其概念范畴与任务定义，并与其他相似的属性识别任务进行对比。其次，本文对目前主流的行人属性识别数据库进行了简单介绍，并从图片和标注两个角度分析了不同数据库之间的异同。再次，本文对深度学习时代以来所提出的各种行人属性识别方法进行了归纳和总结，综述了目前行人属性识别领域的研究现状。最后，本文对监控场景中的行人属性识别存在的问题进行了思考和讨论，并对未来的发展趋势进行了展望。

关键词 计算机视觉, 深度学习, 行人属性识别, 多标签分类, 场景理解。

中图法分类号：TP18

Pedestrian Attribute Recognition in Surveillance Scenes: a Survey

Jia Jian [1),2)] Chen Xiaotang [1),2)] Huang Kaiqi [1),2),3)]

[1)](School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049)

[2)](CRISE, Institute of Automation, Chinese Academy of Sciences, Beijing 100190)

[3)](CAS Center for Excellence in Brain Science and Intelligence Technology, Shanghai 200031)

Abstract The task of pedestrian attribute recognition in surveillance scenes aims to predict the attribute categories of pedestrian images captured by video cameras in surveillance scenes. Due to the complexity of the surveillance scene environment and the fine-grained labeling of pedestrian attributes, the task of pedestrian attribute recognition in surveillance scenes is extremely challenging and has received extensive attention from the industry and academia. This paper reviews the research progress of pedestrian attribute recognition in surveillance scenes. Firstly, we give its conceptual scope and task definition and compare it with similar attribute recognition tasks. Secondly, this paper briefly introduces the mainstream pedestrian attribute recognition datasets and analyzes the similarities and differences between different datasets from picture and annotation perspectives. Again, this paper summarizes and concludes the various pedestrian attribute recognition methods proposed since the era of deep learning and reviews the current research status in pedestrian attribute recognition. Finally, this paper considers and discusses the problems of pedestrian attribute recognition in surveillance scenes and provides an outlook on the future development trend.

Key words computer vision, deep learning, pedestrian attribute recognition, multi-label classification, scene

---

本课题得到国家自然科学基金(No.61721004,No.61876181)、中国科学院项目(No. QYZDB-SSW-JSC006)、中国科学院战略性先导科技专项资助(No.XDA27000000)以及中国科学院青年创新促进会资助.**贾 健**，博士研究生，主要研究领域为计算机视觉、模式识别、行人属性识别. E-mail: jiajian2018@ia.ac.cn. **陈晓棠**，博士，副研究员，主要研究方向为计算机视觉、模式识别.E-mail: xtchen@nlpr.ia.ac.cn. **黄凯奇**（**通信作者**），博士，研究员，计算机学会(CCF)杰出会员,主要研究领域为计算机视觉、模式识别、视觉监控，认知决策等.E-mail: kqhuang@nlpr.ia.ac.cn.

understanding

1 引言

随着智慧城市的发展和公共安全需求的增长，智能视频监控摄像头的数量日益增多，通过人工视频监控的方式处理上亿监控摄像头产生的数据已经远远不能满足社会的需要，因此智能视频监控技术[1]应运而生。其中，监控场景中的行人属性识别技术作为智能视频监控技术的一个重要组成部分，由于其在各个任务场景中的广泛应用受到人们的关注。目前，该技术已被广泛应用于智慧安防、广告营销，与商业零售等领域。在智慧安防领域，借助行人属性识别技术，对监控摄像头捕捉的视频及图像进行结构化解析，通过对目标行人进行属性分类或者利用目标属性进行行人检索，来获得相应的属性以及行人结果，为公共安全提供信息。在广告营销领域，通过对目标场景中出现的行人属性进行实时分析，来判断该场景中行人的消费需求并据此投放相匹配的广告信息，实现广告的精准投放。在商业零售领域，运用行人属性识别技术，分析入店客户的属性与行为信息，进而开展针对性的服务。因此，如何能够在复杂的监控场景中高效的识别目标行人属性，提高行人属性识别技术的性能，成为了工业界和学术界的研究重点。

监控场景中的行人属性识别任务，顾名思义，旨在为监控场景中视频摄像头捕捉的行人图片检测其属性类别，例如性别，年龄，服饰，配饰等等。作为多标签分类的子任务，不同于一般单标签分类任务（多类别分类任务）以及通用的多标签分类任务，行人属性识别任务在图片内容以及标签类别等两方面都有区别于其他相似任务的本质特征。在任务的图片内容方面，如图 3 所示，行人属性识别任务的图片首先采集自监控摄像头拍摄的视频帧，并通过行人检测器(pedestrian detector)或者手工裁剪来获得行人边界框(bounding box),最后将以行人为中心的检测框图片作为最终的行人图片。在任务的标签类别方面，行人属性由多个可见的，客观的，细粒度的语义类别组成。其中，行人属性的可见性是指该属性在图片中是可观测到的，该特点是监控场景中的行人属性识别任务（pedestrian attribute recognition in surveillance）区别于面向行人再识别（person re-identification）的行人属性识别（ReID-oriented pedestrian attribute recognition）任务的本质特征。行人属性的客观性是指该属性是客观存在的，而不依赖与人类的主观判断。例如，性别，年龄，衣着等属性是客观的，而美丑，善恶等属性则是主观的。行人属性的细粒度是指属性标签之间的差异度较小，该特点是行人属性识别任务区别于多标签分类任务的显著特征。例如，相比多标签分类任务中的“person”,“dog”,“car”等类别之间的差异,“ub-Shirt”,“ub-TShirt”,“ub-Cotton”，“ub-ShortSleeve”等行人属性之间的差异更小，并导致了噪声标注以及语义模糊的问题[2]。

由于监控场景中的复杂环境以及细粒度的属性标签，监控场景中的行人属性识别任务具有如下挑战。首先，行人图片采集自视频监控摄像头的视频帧序列，并经过裁剪得到，因此行人图片的分辨率较低，大多集中在 200x100 的分辨率附近。其次，受限于监控摄像头的部署位置以及监控时间，导致行人图片中光照差异较大并且行人之间的姿态变化显著。再次，集中在同一人体部件的行人标签表观差异小，语义特征较为接近，因此容易造成标注错误。最后，行人属性之间的关系复杂，既存在概率关系，例如“Male”和“Bald”（男性有一定概率秃顶）,也存在因果关系，例如“Dress”和“Female”（裙子可以推断出是女性。同时，目前监控场景中的行人属性识别任务仍局限在同一域（domain）内进行训练和测试，并没有涉及跨域（cross-domain）方面的内容。监控场景中的行人属性识别任务除了单独作为一项任务之外，也可以用来辅助行人再识别[3]，行人检索[4]，行人跟踪，场景分析等领域。因此，监控场景中的行人属性识别任务在实际应用和学术研究中都有着较高的价值并受到人们的广泛关注。

与大多数计算机视觉基础任务一样，例如图像分类[5]，物体检测和分割[6]，目标跟踪[7],行为识别[8]等，深度学习[9]的快速发展极大地推动了监控场景中的行人属性识别技术的进步。监控场景下的行人属性识别任务在 2013 年首次由 Zhu 等人[10]提出，并在 2015 年由 Li 等人[11]首次将深度学习引入其中。迄今为止，各种数据库[4,12-15]和方法[16-21]被不断提出，进一步提高了行人属性的识别性能，完善了行人属性识别领域的发展。然而，现有工作并没有对监控场景下的行人属性识别任务给出明确的定义，并导致大量的工作将监控场景中的行人属性识别任务(pedestrian attribute recognition

in surveillance)与其他相似任务相混淆，例如人体属性识别(human attribute recognition)以及面向行人再识别的行人属性识别(ReID-oriented pedestrian attribute recognition)。同时，由于缺少完整准确的监控场景中的行人属性识别任务的定义，以及缺乏对该任务在实际应用场景中本质要求的理解，导致现有的部分监控场景中的行人属性数据库存在“数据泄露”的问题，因此现有工作提出的各种算法在部分数据库上的性能并不能反映模型真实的泛化性能，误导了监控场景中的行人属性识别领域的发展。

因此，本文从五个方面梳理总结了监控场景中的行人属性识别任务，探讨了该任务未来的发展趋势。第一，本文从监控场景中行人属性识别被提出的背景出发，给出该任务完整精确的定义。本文认为作为产生于工业界并有着广泛实际应用前景的任务，行人属性识别的定义必须考虑实际应用场景的要求。同时，作为受到广泛关注和研究的学术任务，行人属性识别的定义也要考虑学术规范和准则。第二，本文在给定监控场景中行人属性识别任务定义的基础上，将监控场景中的行人属性识别任务，面向行人再识别的行人属性识别任务以及人体属性识别任务（human attribute recognition）三个任务进行对比，从数据采集方式，属性标注准则两个主要方面，分析总结出行人属性识别区别于其他任务的本质要求与特点。第三，在给定准确完整定义的基础上，本文对目前学术界主流的多个监控场景中的行人属性数据库进行了分析。数据库是衡量算法有效性的基础，一个合理的数据库构建决定了算法性能的可靠性。然而，通过对现有数据库划分标准的分析，本文发现部分行人属性数据库并不能准确衡量模型的泛化性能。第四，本文以监控场景中的行人属性识别方法提出的时间为主线，对现有算法进行介绍和对比，并归纳总结出各自的特点。第五，通过梳理近年来监控场景中的行人属性识别技术的发展脉络，文本对该领域存在的困难和挑战进行了阐述并对监控场景中的行人属性识别的发展方向进行了思考，探讨了这一领域未来的研究方向。

2 监控场景中的行人属性识别的定义

作为多标签分类（general multi-label classification）的子任务，监控场景中的行人属性识别任务与多标签分类任务有密不可分的联系，因此，本节首先从概念归属的角度出发，将监控场景中的行人属性识别任务与多标签分类任务进行对比，对监控场景中的行人属性识别任务的概念范畴进行界定，如图 1 所示。其次，相比于属性识别任务中的人体属性识别（human attribute recognition）[22]以面向行人再识别的行人属性识别任务（ReID-oriented pedestrian attribute recognition），监控场景中的行人属性识别任务又有自身的特点和要求。同时，为了解决目前行人属性识别在不同文献[23-26]中的称谓混淆或指代内容不一致的问题，本节从图片采集方式以及属性标注准则两个方面，厘清了上述三个任务之间的差异，并给出了监控场景下的行人属性识别任务的完整精确的定义。

2.1 任务范畴

多标签分类任务（general multi-label classification）作为目前覆盖内容最广的分类任务，[27]旨在为存在多个不同类别物体的图片正确预测多个物体类别标签。常用的多标签分类任务的数据库有 COCO[28]，PASCAL-VOC[29]，以及 OpenImagev4[30]等。其中，单标签分类任务，即多类别分类任务，是多标签分类任务中，每张图片标签数量为 1 的特例，例如 ImageNet[31]，CIFAR10[32],MINIST[33]。属性识别任务作为多标签分类任务中重要的一个组成部分，主要包括人脸属性识别任务（face attribute recognition），服饰属性识别任务（fashion attribute recognition），人体属性识别任务（human attribute recognition）以及本文重点研究的行人属性识别任务（pedestrian attribute recognition）。而行人属性识别任务主要由监控场景中的行人属性识别任务以及面向行人再识别的属性识别任务构成。各个任务的从属关系如图 1 所示。

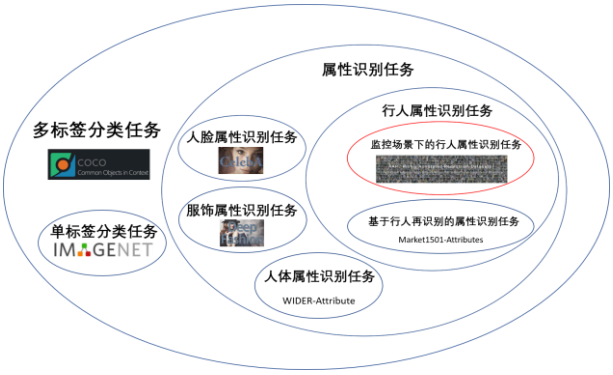

图 1 监控场景中的行人属性识别任务的概念范畴。监控场景中的行人属性识别任务按照从属关系，分别属于多标签分类任务（general multi-label classification），属性识别任务

(attribute recognition)，以及行人属性识别任务(pedestrian attribute recognition)。

多标签分类任务（general multi-label classification）或一般多标签分类任务的图片内容丰富，环境变化多样，不局限于单一的场景。从图片角度来说，如图 2 和图 3 所示，多标签分类数据库[28-30]的图片大多来源于搜索引擎或采集自专业摄像机的拍摄图片，由于光照充足，相机离目标拍摄物体较近，因此图片分辨率高，大多图片的分辨率集中在 600x400 附近，图片中的各类物体清晰。相反，如图 2 和图 3 所示，监控场景中的行人图片[4,10,12-14]均采集自远距离部署的视频监控摄像头，并经过行人检测器（pedestrian detector）或手工裁剪而成，因此图片分辨率较低，加之不同图片间光照差异显著，导致行人属性较为模糊。综上，行人属性数据库的图片质量远低于多标签分类数据库。除图片质量之外，多标签分类任务与监控场景中的行人属性识别任务的不同之处还在于标签类别。多标签分类任务中的图片标签是粗粒度的物体类别，例如人，狗，自行车等，这些类别之间语义差别明显，表观差异显著。然而，监控场景中的行人属性识别任务中的图片标签是细粒度的属性类别，例如 T 恤，POLO 衫，短袖等，这些属性标签本身语义差距较小，再结合行人图片较为模糊的问题，因此不同属性标签之间的表观差异较小，造成属性标签之间的语义模糊。例如 RAPv1[12]数据库中“lb-LongTrousers”（紧身裤），“lb-Jeans”（牛仔裤），“lb-TightTrousers”（长裤）等三个属性。并且由于属性标签的语义模糊问题，容易导致错标和漏标等问题。因此与多标签任务的物体类别相比，行人属性识别的属性标签属于更细粒度的标注，标签之间差异更小，且存在较多噪声。

图 2 多标签分类任务和行人属性识别任务在图片方面的比较。我们以最常见的 COCO[28]数据库为例来代表多标签分类任务。监控场景中的行人属性识别任务以目前规模最大的数据库 RAPv2[4]以及 PA100k[14]为例。

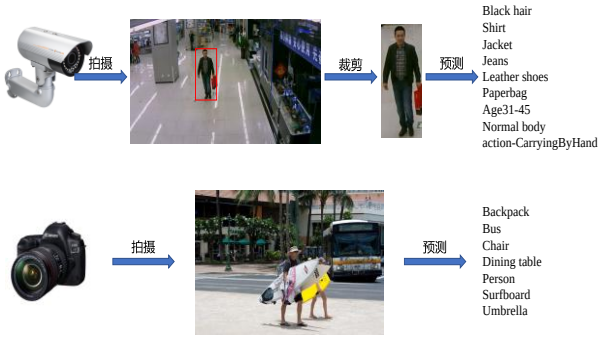

图 3 多标签分类任务和行人属性识别任务的流程。

监控场景中的行人属性识别任务首先由 Zhu 等人在 APiS（Attributed Pedestrians in Surveillance）数据库[10]中提出，并分别在 2014 年 Deng 等人提出的 PETA,2016 年 Li 等人提出的 RAPv1,2017 年 Liu 等人提出的 PA100k,以及 2016 年 Li 等人在 RAPv1 基础上扩展得到的 RAPv2 等数据库中得到完善和发展。在这个过程中，行人属性数据库 PARSE-27k（Pedestrian Attribute Recognition on SEquences）由 Sudowe 等人于 2015 年提出，该数据库是由一个移动的摄像机拍摄，虽然拍摄的对象是城市街道中的行人，但由于图片并非采集自监控场景中，因此该数据库对应的任务不能称为监控场景中的行人属性识别任务。

2.2 监控场景中的行人属性识别与其他相似任务的对比

尽管监控场景中的行人属性识别技术在深度学习时代有了极大的发展和提高，但目前已有的文献并没有给出监控场景中行人属性识别任务明确完整的定义。因此，不少工作将监控场景中的行人属性识别任务与人体属性识别以及面向行人再识别的行人属性识别任务相混淆，将针对不同任务的数据库或方法混为一谈。例如，综述[25]中，将人体属性识别的数据库 WIDER 归为监控场景中的行人属性识别数据库；综述[24]以及方法[34,35]中忽略了面向行人再识别的行人属性识别任务和监控场景中的行人属性识别任务的差异，将其视为同类；方法[23,36]则将人体属性识别和监控场景中的行人属性识别任务视为一种任务。因此本节首先从监控场景中的行人属性识别任务提出的背景出发，给出该任务的完整定义，并厘清与人体属性识别以及面向行人再识别的行人属性识别这两个任务的区别。

从实际应用角度来说，监控场景中的行人属性

识别任务既可以作为独立任务，对监控场景下的行人图片预测其属性类别，也可以将预测出的属性作为语义信息辅助行人再识别，行人跟踪等任务。不论是作为独立任务还是辅助任务，监控场景中的行人属性识别的最终目的都是对在训练过程中未见过的行人及图片预测出正确的属性，其核心要求都是要在训练过程中未见过的行人图片上表现出良好的泛化性能。因此本文采用监控场景中的行人属性识别任务的定义[15]如下:

定义: 给定包含监控摄像头捕捉的行人图片以及对应的属性标签的训练集 $D=\{x(y_i\ i=\ ).,.N$ ，行人属性识别旨在对测试集的行人图片 x_i 预测多个属性标签 $y_i \in \{0,1\}^M$ ，其中测试集的行人身份 I_{test}（pedestrian identity，以下简称行人）从未出现在训练集中，即训练集和测试集的行人符合零样本的设定 $I_{train} \cap I_{test} = \varnothing$。

与之前文献中提及的行人属性识别任务不同，本文采用的行人属性识别任务的定义，其核心的要求不仅仅体现在正确预测行人图片中出现的属性类别，同时要求对训练集未出现过的测试集中的行人的图片进行正确的属性预测。根据定义，行人属性识别任务的本质特征是：在一部分行人 I_{train} 的图片上训练得到的模型要在另一部分行人 I_{test} 的图片上有良好的泛化性能。具体来说，即训练集的行人身份集合与测试集的行人身份集合之间没有重叠 $I_{train} \cap I_{test} = \varnothing$，而这一要求被之前的相关工作[4, 12,13]所忽略。同时，这一要求也是行人属性数据库能够正确衡量各种行人属性识别方法泛化性能的基础。

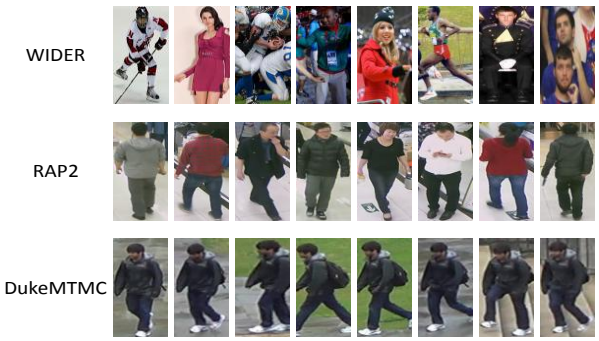

图 4 监控场景中行人属性识别数据库,人体属性数据库，以及面向行人再识别的行人属性识别数据库的对比。其中人体属性数据库以 WIDER 为代表。监控场景中行人属性识别数据库以 RAPv2 为代表。面向行人再识别的行人属性识别数据库以 DukeMTMC 为代表

图 5 面向行人再识别的行人属性识别和监控场景中的行人属性识别之间的属性标注差异。其中面向行人再识别的行人属性识别以 DukeMTMC 数据库为代表，监控场景中的行人属性识别以 RAP 数据库为代表。

与行人属性识别任务相似的任务主要有两个，分别是人体属性识别任务（human attribute recognition）[22,37,38]以及面向行人再识别的行人属性识别任务 (ReID-oriented pedestrian attribute recognition) [3]。人体属性识别任务分别由 Bourdev 等人在 BAP (Berkeley Attributes of People) [37] 数据库以及 Sharma 等人在 HAT (Human ATtributes) [38] 数据库中提出。受到 HAT 和 BAP 数据库的启发，同时为了解决之前数据库的局限性，Li 等人则提出了一个大规模人体属性识别数据库 WIDER [22]。尽管 BAP，HAT，以及 WIDER 等数据库在图片数量以及属性标注类别上存在诸多差异，但是这三个数据库的图片均采集自搜索引擎，其属性类别均包括性别，服饰，配饰等人体基本属性类别。综合考虑上述三个典型的人体属性识别数据库中的共同之处，本文认为人体属性识别任务旨在为通用场景中以人为中心的图片进行多个人体属性的预测。与监控场景下的行人属性识别的图片不同，人体属性识别的图片由于来自搜索引擎，场景非常丰富，人体姿态多样，同一张图片往往包含多个人物，并且基本不存在相同人物的相似图片，如图 4 所示。而监控场景中的行人属性识别的图片则来自监控场景中的视频摄像头，人体姿态均以站立为主，一张图片中大多只包含一个行人，并且由于目前行人属性识别数据库大多采集自同一场景中的不同摄像头，因此图片背景较为单一和

相似。另外，监控场景中的行人属性识别技术，可用于辅助行人跟踪，行人检测等监控场景中的其他任务，而人体属性识别任务并不具备这一特征。

面向行人再识别的行人属性识别任务则由Lin[3]等人为了辅助行人再识别任务，提高行人再识别的性能而提出。该任务的数据库是在已有的行人再识别数据库 Market[39]和 DukeMTMC[40]的基础上，对图片按照行人级别的标注而构成。尽管该任务在图片采集方面与监控场景中的行人属性识别任务比较类似，但是两者在图片标注方面存在着显著差异。如图 5 所示，以 DukeMTMC 为代表的面向行人再识别的行人属性识别数据库中的属性采用行人身份（pedestrianidentity）级别的标注，即同一行人的所有图片均含有相同的属性标注，而不论该属性是否出现在图片中。以“Bag”和“HandBag”两个典型的属性为例，在部分行人图片中由于遮挡导致该属性不可见，但是面向行人再识别的行人属性识别数据库中这些图片依然被标注为该属性的正样本。相反，在以 RAPv2 为代表的监控场景中的行人属性识别数据库中的属性采用图片实例级别的标注，即所有图片均按照属性是否出现来进行标注。以“shoes-Leather”，“action-Holding”和“attachment-Box”属性为例，对于同一行人的不同图片，由于部分图片中目标属性不存在，因此，部分图片为目标属性的正样本（绿色边界框），部分图片为目标属性的负样本（红色边界框）。因此，由于面向行人再识别的行人属性识别数据库在图片标注方式上的缺陷，导致该任务对于属性识别任务来说存在大量的噪声样本。另外，由于面向行人再识别的行人属性识别数据库采集的时间较为集中，因此在服饰等属性中出现了严重的类别不均衡现象，例如 Market1501 数据库采集自夏季， 因此数据库中存在大量“短袖”属性的正样本而几乎没有“长袖”属性的正样本。综上，不论是从任务提出的背景，还是图片标注准则等方面考虑，面向行人再识别的行人属性识别与监控场景中的行人属性识别任务均存在较大差异，并不能归为同类的任务。 三种任务之间的对比如表 1 所示。

表 1 监控场景中的行人属性识别，面向行人再识别的行人属性识别以及人体属性识别任务之间

任务	图片采集方式	数据标注方式	应用场景	任务目的	主要难点
监控场景中的行人属性识别	视频监控摄像头	图片实例级别	监控场景	属性分类	较低的图片分辨率，显著的光照差异
人体属性识别	搜索引擎	图片实例级别	日常生活场景	属性分类	显著的人体姿态差异，复杂的拍摄现场
面向行人再识别的行人属性识别	视频监控摄像头	行人身份级别	监控场景	辅助行人再识别	将局部特征与全局特征相融合

3 监控场景中的行人属性识别数据库及评价指标

本节主要介绍了监控场景中的行人属性数据库以及相应的评价指标。同时，为了与面向行人再识别的行人属性识别任务和人体属性识别任务从数据库的层面进行区分，本节也对这两个任务的数据库进行了简单介绍。三种任务数据库之间的信息对比如表 1 所示。

3.1 监控场景中的行人属性识别数据库

监控场景中的行人属性识别任务首先由 APiS 数据库提出，并在 PETA[13],RAP1[12],PA100K[14]，以及 RAP2[4]等数据库中得到进一步发展。

APiS(Attributed Pedestrians inSurveillance)数据库由 Zhu[10]等人在 2013 年提出。该数据库包含 3661 张行人图片，每张行人图片标注了 11 个二元属性以及 2 个多类别属性，例如性别，服饰，背包，以及衣着颜色等。为了构建一个包含多个场景的行人属性识别数据库，该数据库收集了来自四个数据库的行人图片，分别是 KITTI [41]数据库, CBCLSS [42]数据库, INRIA [43]数据库以及由部署在火车站的监控摄像头采集的 SVS 数据库。

PETA (PEdesTrian Attribute) 数据库由 Deng [13]等人在 2014 年提出。该数据库包含 19000 张行人图片，图片的分辨率从 17 × 39 到 169 × 365 像素，同时每张行人图片标注了 61 个二元属性和 4 个多类别属性，例如年龄，性别，服饰，配饰等。该数据库作为第一个大规模的行人属性数据库，其图片收集自 10 个小规模的行人再识别（person re-identification）数据库，因此图片之间存在拍摄角度，光照，场景信息等因素的显著差异。该数据

库将 19000 张图片随机划分为包含 9500 张图片的训练集，1900 张图片的验证集，以及 7600 张图片的测试集。尽管该数据库标注了超过 60 个属性，但是由于属性分布的不均衡，只有 35 个属性被选择用于训练以及测试。值得注意的是，该数据库的标注采用的是行人身份级别(identity-level)的标注，即，同一个行人的不同图片拥有完全一样的属性标注。因此，当某些属性由于人体姿态变化出现遮挡而不可见时，这种属性标注的方式会引入噪声，从而影响方法的性能。

表 2 三种不同任务数据库的对比。其中属性数量是指用于模型训练和测试的属性数量，而非数据库标注的属性数量。“--” 表示无法统计或原始文献未给出的结果。N_{all}，N_{train}，N_{valid}，N_{test} 分别表示数据库中所有样本的数量，以及训练集，验证集和测试集的样本数量。场景中的“混合监控”是指室内以及室外监控

任务	数据库	属性数量	N_{all}	N_{train}	N_{valid}	N_{test}	摄像头数量	行人数量	标注准则	场景
监控场景中的行人属性识别	APiS	13	3661	-	-	-	-	-	实例级别	混合监控
	PETA	35	19,000	9,500	1,900	7,600	-	8,699	实例级别	混合监控
	RAPv1	51	41,585	33,268	-	8,317	23	-	实例级别	室内监控
	RAPv2	54	84,928	50,957	16,986	16,985	25	2,589	实例级别	室内监控
	$PETA_{ZS}$	35	19,000	11,241	3,826	3,933	-	8,699	实例级别	混合监控
	RAP_{zs}	53	26,632	17,062	4,648	4,928	25	2,589	实例级别	室内监控
	PA100K	26	100,000	80,000	10,000	10,000	-	-	实例级别	室内监控
人体属性识别	BAP	9	8,035	2,003	2,010	2,010	4,022	-	实例级别	日常生活
	HAT	27	9,344	3,500	3,500	3,500	2,344	-	实例级别	日常生活
	WIDER	14	13,789	5,509	5,509	1,362	6,918	-	实例级别	日常生活
面向行人再识别的行人属性识别	Market1501	27	32,668	12,936	12,936	-	19,732	6	行人身份级别	室外监控
	DukeMTMC	23	34,183	16,522	16,522	-	17,661	8	行人身份级别	室外监控

RAP1(Richly Annotated Pedestrian)数据库由 Li[12]等人在 2016 年提出。为了解决 PETA 数据库中图像分辨率低，缺少人体视角以及人体部件标注等问题，RAP1 数据库收集了 41585 张图片，并对每张图片标注了 69 个二元属性和 3 个多类别属性。与 PETA 数据库采用行人身份级别的标注不同，RAP1 数据库采用了图片实例级别（instance-level）的标注，即根据每张图片中行人实际存在的属性进行标注，从而避免了引入噪声的风险。除了常规的服饰，年龄等属性之外，RAP1 数据库还为行人图片标注了人体视角，遮挡类型，人体部件位置三种信息。该数据库将 41585 张图片随机划分为包含 33268 张图片的训练集以及 8317 张图片的测试集。该数据库采集自一个室内商场的 26 个监控摄像头。

PA100K(Pedestrian Attribute-100K)数据库由 Liu[14]等人在 2017 年提出，是目前为止最大的监控场景下行人属性识别数据库。该数据库拥有 100000 张行人图片，包括 80000 张训练集图片，10000 张验证集和 10000 张测试集图片，每张行人图片被标注了 26 个属性。与 PETA 和 RAP1 数据库采用的将图片随机划分为训练集和测试集的标准不同，PA100K 按照行人的身份标注划分训练集和测试集，使训练集和测试集的行人身份符合零样本的设定，即，测试集的行人身份没有出现在训练集中。

RAP2(Richly Annotated Pedestrian)数据库由 Li[4]等人在 RAP1 的基础上拓展而成，该数据库主要用于行人属性识别和行人检索。该数据库共包含 2589 个行人的 84928 张图片，每张图片共标注了 72 个属性。除常见的性别，年龄，服饰等属性之外，RAP2 数据库还增加了姿态，动作等属性标签。该数据库将 84928 张图片分为 50957 张训练集图片，16986 张验证集图片，以及 16985 张测试集图片。尽管 RAP2 相比 RAP1 有更多的训练样本，但是目前学术界主流的方法依然采用 RAP1 数据库来进行性能评测。该数据库采集自一个室内商场的 25 个监控摄像头。

$PETA_{ZS}$ 数据库由 Jia[15]等人在 PETA 数据库的基础上重新划分而成。该数据库解决了原始 PETA

数据库中训练集和测试集存在的行人身份重叠的问题，避免了数据泄露。该数据库将 19000 张图片重新根据行人身份零样本的设定划分为包含 5211 个行人，11051 张图片的训练集，包含 1703 个行人，3980 张图片的验证集，以及 1785 个行人，3,969 张图片的测试集。数据库图片的标注与原数据库保持一致。

RAP_{ZS} 数据库由 Jia[15]等人在 RAP2 数据库的基础上重新划分而成。该数据库从 RAP2 数据库中挑选出含有行人身份标注的 26632 张图片，并按照行人身份零样本的设定划分训练集和测试集，其中训练集包含 1509 个行人的 14729 张图片，验证集包含 546 个行人的 5961 张图片，测试集包含 535 个行人的 5948 张图片。数据库图片的标注与原数据库保持一致。

3.2 人体属性数据库

目前常用的人体属性数据库主要有三个，分别为 BAP[37],HAT[38]以及 WIDER[22].

BAP(Berkeley Attributes of People)数据库由 Bourdev 等人在 2011 年提出。该数据库包含 8035 张以人为中心的图片，包括 2003 张训练集图片，2010 张验证集图片，以及 4022 张测试集图片。这些图片分别采集自 H3D[44] 和 PASCAL VOC2010[29] 数据库的训练集和测试集。与 PASCAL 数据库中低分辨率版本的图片不同，该数据采用了在 Flickr 上收集到的高分辨率的对应图片。对于图片中的人物，我们将高分辨率图片沿着人物的边界进行裁剪，使其在可见的范围内保留足够的背景，并将图片进行缩放，使臀部和肩膀之间的距离为 200 像素。该数据库使用亚马逊 MechanicalTurk[45]平台，通过五位独立的数据标注员对所有九个属性进行标注。这些属性包括性别，发型，服饰等。

HAT(Human ATtributes)数据库由 Sharma 等人在 2011 年提出。该数据库包含 9344 张图片，每张图片标注了 27 个属性,例如年龄，服饰，以及配饰等。HAT 数据库中的图片均采集自 Flickr 网站，作者通过输入超过 320 个索引信息来搜索与索引相似度最高的图片。并利用人体检测器（person detector）[46]从搜索到的图片中裁剪出人体图片。该数据库包含 3500 张训练集图片,3500 张验证集图片,以及 2344 张测试集图片。

WIDER-Attribute 数据库由 Li[22]等人在 2016 年提出。该数据库包含 13789 张图片，每张图片标注了 14 个人体属性，这些属性主要由年龄，服饰，配饰组成。这些图片均采集自 Xiong 等人提出的 WIDER（Web Image Dataset for Event Recognition）图片数据库[47]。在数据标注过程中，作者首先对图片中出现的人体标注边界框（bounding box），并对每个边界框中的人体标注 14 个属性。该数据库共包含 5509 张训练图片、1362 张验证图片以及 6918 张测试图片。同时为了利用场景的上下文信息，作者将数据库中的所有图片划分为 30 个场景类别。

通过对三个人体属性数据库进行分析可以发现，目前主流的人体属性数据库图片都采集自搜索引擎，例如 BAP 以及 HAT 的主要图片采集自 Flickr，而 WIDER-Attribute 的图片则来自 Google 以及 Bing 网站。因此，在人体属性识别领域中，人体属性识别数据库的图片分辨率较高，图片中场景复杂，人物数量多，属性与场景有着密切的联系。并且，由于采集方式的限制，训练集和测试集不会出现相似度较高的图片，也不会出现属于同一人物的图片。模型的泛化性能够很好的在测试集上得到验证。

3.3 面向行人再识别的行人属性识别数据库

为了辅助行人再识别任务，Lin[3]对两个主流的行人再识别数据库 Market1501[39] 和 DukeMTMC[40]的图片进行了属性标注。与人体属性识别数据库以及监控场景中的行人属性识别数据库不同，面向行人再识别的行人属性识别数据库均采用了行人身份级别(identity-level)的标注，并且其属性类别主要是颜色属性为主。另外，由于 Market1501 数据库和 DukeMTMC 数据库分别采集自夏天和冬天的室外场景，因此其服饰属性存在严重的分布不均衡现象。

Market1501-Attribute 数据库共包含 1501 个行人的 32668 张图片，其中训练集包含 751 个行人的 12936 张图片，测试集包含 750 个行人的 19732 张图片。每张图片标注了 27 个属性，其中 15 个为颜色属性。该数据库中的图片采集自部署在校园超市的 6 个监控摄像头。

DukeMTMC-Attribute 数据库共包含 1812 个行人的 34183 张图片，其中训练集包含 702 个行人的 16522 张图片，测试集包含 1110 个行人的 17661 张图片。每张图片标注了 23 个属性，其中 15 个为颜色属性。该数据库中的图片采集自杜克大学（Duke University）的 8 个监控摄像头。

3.4 性能评价指标

监控场景中的行人属性识别任务主要采用的指标可以分类两大类，分别为属性级别（attribute-level）的指标以及样本实例级别（instance-level）的指标。其中属性级别的指标为平均准确率（mean accuracy，mA），该指标综合衡量了测试方法在属性层面的性能。该指标通过分别计算目标属性正负样本的召回率再取平均得到，其计算公式如下：

$$mA = \frac{1}{M}\sum_{j=1}^{M}\frac{1}{2}\left(\frac{TP^j}{TP^j+FN^j}+\frac{TN^j}{TN^j+FP^j}\right)$$

其中 TP^j, TN^j, FP^j, FN^j 分别是第 j 个属性的真阳性（true positive），真阴性（true negative），假阳性（false positive），假阴性（false negative）的样本数量，M 则代表数据库中所有属性的数量。样本实例级别的指标共有四个，分别为准确率（accuracy，Acc），精确率(precision, Prec)，召回率(recall, Recall)以及 F1 值。各个指标的计算公式如下：

$$Acc = \frac{1}{N}\sum_{i=1}^{N}\frac{TP_i}{TP_i+FP_i+FN_i}$$

$$Prec = \frac{1}{N}\sum_{i=1}^{N}\frac{TP_i}{TP_i+FP_i}$$

$$Recall = \frac{1}{N}\sum_{i=1}^{N}\frac{TP_i}{TP_i+FN_i}$$

$$F1 = \frac{1}{N}\sum_{i=1}^{N}\frac{2\cdot Prec\cdot Recall}{Prec+Recall}$$

其中 TP_i, FP_i, FN_i 分别是第 i 个样本的真阳性（true positive）,假阳性（false positive）,以及假阴性（false negative）的属性数量，N 代表所有样本的数量。

人体属性识别任务主要采用的指标为全类别平均精确率(mean Average Precision, mAP)。全类别平均正确率 mAP 首先计算各个属性类别上的平均精确率（average precision，AP），再在所有属性上取平均所得到。其中平均精确率 AP 的计算分为两步，对于目标属性的所有样本，首先根据模型预测得到的置信度，将所有样本按照置信度升序排列，并计算该属性在累加样本上的召回率（Recall）和精确率（Precision），并以召回率为横轴，精确率为纵轴，画出 Precision-Recall（PR）曲线的联系再取 PR 曲线下面积作为 AP 值。

面向行人再识别的行人属性识别任务则主要采用全类别平均精确率 mAP 以及四个实例级别的指标 Acc,Prec,Recall,F1，其含义和计算方法与监控场景中的行人属性识别中的评价指标相同。

3.5 行人属性数据库的总结和对比

通过对三种属性识别任务的数据库的分析可知，由于图片的采集方式不同（来自搜索引擎或者来自监控场景的摄像头）以及属性标注方式不同（行人身份级别标注和图片实例级别标注），三种识别任务的数据库有各自的特点和适用场景， 如表 2 所示。监控场景中的行人属性识别的图片大多采集自视频监控摄像头。其原始图片通过行人检测器或者手工标注裁剪出以人为中心的行人边界框图片。因此，行人属性识别数据库图片大多只包含一个行人，并拥有完整的行人部件。但由于设备限制以及监控摄像头与拍摄的行人位置较远，导致图片分辨率低，增大了属性识别的难度。同时，数据库中的图片是在短时间内由连续的多张图片帧裁减而成，例如 RAP,RAP2 数据库的图片是以每秒 15 帧的速率采集得到，导致同一行人的图片之间变化较小，相似度较高，并由于数据库训练集和测试集的随机划分，造成了训练集和测试集有大量相似的图片。因此，模型的泛化性能在 RAP,RAP1,PETA 上不能得到充分的衡量。人体属性识别数据库的图片采集自搜索引擎，因此图片分辨率高，图片中人物数量较多，同时图片中大多数人体部件不完整。人体属性数据库主要用于研究开放场景（unconstrained settings）中的属性识别[22]，同时由于其图片均采集自搜索引擎，所以其训练集和测试集的图片不会出现相同行人的相似度很高的图片，模型的泛化性能能够得到充分的衡量。面向行人再识别的行人属性识别数据库主要用于辅助行人再识别任务，其图片同样采集自监控摄像头，但是图片标注方法采用基于行人身份级别的标注，因此数据库中存在大量的噪声样本。除此之外，与监控场景中的行人属性识别数据库筛选出样本分布较为均衡的属性作为训练和测试属性不同，面向行人再识别的行人属性识别数据库将所有标注属性用于训练和测试，因此数据库存在显著的样本分布不均衡问题。

4 基于深度学习的行人属性识别方法

行人属性识别任务自提出以来，由于其广泛的应用前景以及适用于研究诸如长尾分布问题

（long-tailed distribution），特征表达问题（feature representation），度量学习（metric learning）等学术问题而受到工业界和学术界的广泛关注。众多的研究人员为了解决行人属性识别任务中涉及的不同问题而提出各种创新性的方法，显著的提高了行人属性识别的性能，推动并完善了行人属性识别领域的发展。本节主要对深度学习以来行人属性识别领域中监控场景的行人属性识别方法以及面向行人再识别的行人属性识别方法进行简单的介绍和总结。同时，由于人体属性识别任务与行人属性识别任务较为相似，并被一些研究人员混用，因此本节也对人体属性识别的方法进行了简要介绍。

表 3 监控场景中的行人属性识别方法在 PETA，RAPv1，PA100K 三个数据库上的性能。其中 “--”表示原文未展示相关结果

方法	主干网络	PETA					RAPv1					PA100k				
		mA	Accu	Prec	Recall	F1	mA	Accu	Prec	Recall	F1	mA	Accu	Prec	Recall	F1
DeepMAR[11] (*ACPR15*)	CaffeNet	82.89	75.07	83.68	83.14	83.41	73.79	62.02	74.92	76.21	75.56	72.70	70.39	82.24	80.42	81.32
WPAL[48] (*BMVC'17*)	GoogleNet	85.50	76.98	84.07	85.78	84.90	81.25	50.30	57.17	78.39	66.12	–	–	–	–	–
VeSPA[49] (*BMVC'17*)	GoogleNet	83.45	77.73	86.18	84.81	85.49	77.70	67.35	79.51	79.67	79.59	–	–	–	–	–
HPNet[14] (*ICCV'17*)	InceptionNet	81.77	76.13	84.92	83.24	84.07	76.12	65.39	77.33	78.79	78.05	74.21	72.19	82.97	82.09	82.53
JRL[17] (*ICCV'17*) [1]	AlexNet	85.67	–	86.03	85.34	85.42	77.81	–	78.11	78.98	78.58	–	–	–	–	–
JRL[17] (*ICCV'17*)	AlexNet	82.13	–	82.55	82.12	82.02	74.74	–	75.08	74.96	74.62	–	–	–	–	–
LGNet[50] (*BMVC'18*)	Inception-V2	–	–	–	–	–	78.68	68.00	80.36	79.82	80.09	76.96	75.55	86.99	83.17	85.04
PGDM[51] (*ICME'18*)	CaffeNet	82.97	78.08	86.86	84.68	85.76	74.31	64.57	78.86	75.90	77.35	74.95	73.08	84.36	82.24	83.29
GRL[52] (*IJCAI'18*)	Inception-V3	86.70	–	84.34	88.82	86.51	81.20	–	77.70	80.90	79.29	–	–	–	–	–
MsVAA[36] (*ECCV'18*)	ResNet101	84.59	78.56	86.79	86.12	86.46	–	–	–	–	–	–	–	–	–	–
RA[53] (*AAAI'19*)	Inception-V3	86.11	–	84.69	88.51	86.56	81.16	–	79.45	79.23	79.34	–	–	–	–	–
VSGR[54] (*AAAI'19*)[1]	ResNet50	85.21	81.82	88.43	88.42	88.42	77.91	70.04	82.05	80.64	81.34	79.52	80.58	89.40	87.15	88.26
VRKD[55] (*IJCAI'19*)	ResNet50	84.90	80.95	88.37	87.47	87.91	78.30	69.79	82.13	80.35	81.23	77.87	78.49	88.42	86.08	87.24
AAP[56] (*IJCAI'19*)	ResNet50	86.97	79.95	87.58	87.73	87.65	81.42	68.37	81.04	80.27	80.65	80.56	78.30	89.49	84.36	86.85
VAC[27] (*CVPR'19*)	ResNet50	–	–	–	–	–	–	–	–	–	–	79.16	79.44	88.97	86.26	87.59
ALM[19] (*ICCV'19*)	BN-Inception	86.30	79.52	85.65	88.09	86.85	81.87	68.17	74.71	86.48	80.16	80.68	77.08	84.21	88.84	86.46
IA^2□□□[57] (*PRL'19*)	ResNet50	84.13	78.62	85.73	86.07	85.88	77.44	65.75	79.01	77.45	78.03	–	–	–	–	–
JLAC[20] (*AAAI'20*)	BN-Inception	86.96	80.38	87.81	87.09	87.45	83.69	69.15	79.31	82.40	80.82	82.31	79.47	87.45	87.77	87.61
Da-HAR[23] (*AAAI'20*)	ResNet50	–	–	–	–	–	84.28	59.84	66.50	84.13	74.28	–	–	–	–	–
CAS[58] (*ICME'20*)	ResNet34	83.17	78.78	87.49	85.35	86.41	–	–	–	–	–	77.20	78.09	88.46	84.86	86.62
DTM[59] (*Arxiv'20*)	ResNet50	85.79	78.63	85.65	87.17	86.11	82.04	67.42	75.87	84.16	79.80	81.63	77.57	84.27	89.02	86.58
RPAR[15] (*Arxiv'20*)	ResNet50	85.11	79.14	86.99	86.33	86.39	78.48	67.17	82.84	76.25	78.94	79.38	78.56	89.41	84.78	86.55
JLPLS-PAA[34] (*TIP'19*)	SE-BN-InceptionNet	84.88	79.46	87.42	86.33	86.87	81.25	67.91	78.56	81.45	79.98	81.61	78.89	86.83	87.73	87.27
SCRL[35] (*TCSVT'20*)	ResNet50	87.40	–	89.20	87.50	88.30	81.90	–	82.40	81.90	82.10	80.60	–	88.70	84.90	86.80
MTANet[60] (*PRL*)	ResNet152	84.62	78.80	85.67	86.42	86.04	77.62	67.17	79.72	78.44	79.07	–	–	–	–	–

4.1 监控场景中的行人属性识别方法

由于智慧城市概念的普及以及安防摄像头部署数量的增加，行人属性识别逐渐成为了人体属性识别任务中研究最多的任务。本节将以各个方法在 PETA，RAP1，以及 PA100K 三个大规模数据库上的性能为主线，按照提出的时间顺序，对各个方法进行简单的介绍。方法在相关数据库上的性能如表 3 所示。

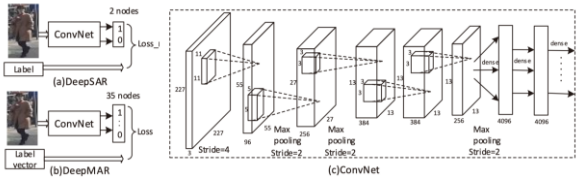

图 6 DeepMAR 方法的模型结构图。

Li 等人提出的 DeepMAR [11] 网络作为第一个将深度学习引入行人属性识别的工作，网络的框架结构如图 6 所示。该方法提出了多属性联合学习框架，并对比分析了单属性学习框架和多属性联合学习框架的性能区别。另外，作者提出了解决样本分布不均衡的加权损失函数。 该损失函数首先计算训练集中每个属性中正样本的比率，将其作为数据库分布的先验， 并结合指数函数用于加强正样本的损失， 使模型的优化方向偏向于真样本。 该加权损失函数作为主流的损失函数， 应用于各种主流的方法中。

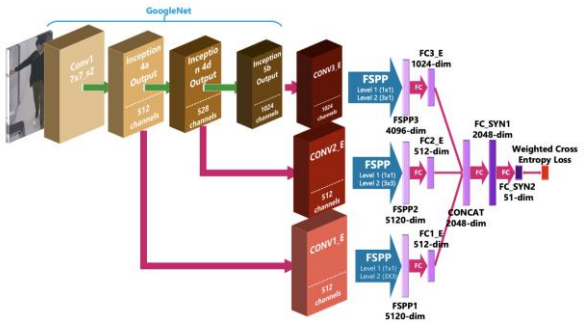

图 7 WPAL 方法的模型结构图

Yu 等人提出基于弱监督的属性识别和定位方法 WPAL [48]，如图 7 所示。该方法借鉴了弱监督检测的思想，将全局最大池化层中的每层通道视为特定物体类别的检测器。该假设认为特征图中最大值对应的空间位置即为目标物体类别所在的位置。该方法以 GoogleNet 作为主干网络， 将检测网络中的空间金字塔池化层（Spatial Pyramid Pooling，SPP） [61] 加以改进并用于主干网络的 Inception4a, Inception4d, 以及 Inception5b 层输出的特征图上。 作者提出了灵活的空间金字塔池化层（Flexible Spatial Pyramid Pooling, FSPP），该网络层扮演了全局最大池化的作用，用于发掘属性相关的特征。 FSPP 是一个两级金字塔结构的网络层， 在第一级结构中对每层特征通道进行全局最大池化，得到每层特征通道的最大响应值。在第二级结构中应用于 Inception4a, Inception4d, 以及 Inception5b 输出特征的 FSPP 分别采用 $3\times 1, 2\times 2$ 和 3×3 的网格进行划分， 在每个网格中采用最大池化得到各个网格的响应，并与第一级结构中得到的特征值进行串接得到输出特征向量。 每一层的特征向量进行降维并串接作为图片的特征向量送入分类层。

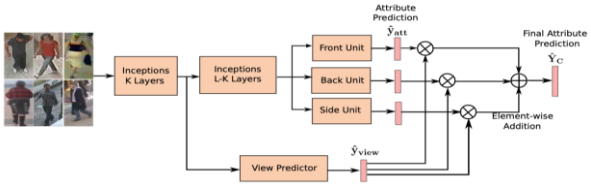

图 8 VeSPA 方法的模型结构图。

Sarfraz 等人利用人体属性与人体视角的强依赖关系作为先验， 将属性识别任务与视角预测任务相结合， 提出多任务学习网络 VeSPA[49]， 如图 8 所示。该网络采用 GoogleNet 作为主干网络，包含一个视角预测分支网络以及三个特定视角下的属性预测分支网络。其中，视角预测网络预测输入图片的视角类别（前视角，背视角，侧视角）的置信度作为加权参数。三个属性预测分支网络则分别预测各个属性类别的概率，再利用视角加权参数进行加权融合，得到最终的属性预测概率。

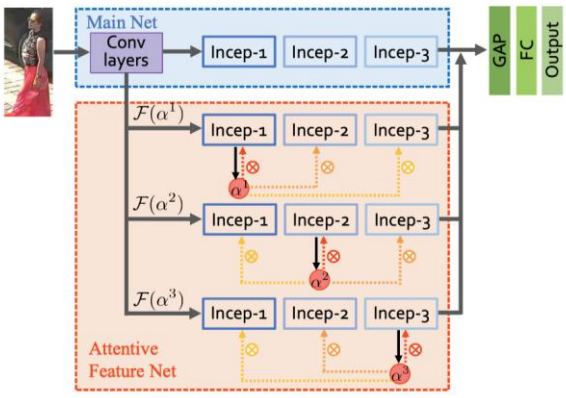

图 9: HPNet 方法的模型结构图。

Liu 等人考虑到属性之间类间差距过小，因此从细粒度识别的角度来解决行人属性识别任务，并提出了 HPNet[14]， 如图 9 所示。作者注意到属性特征不仅需要低级别的特征信息也需要高级别的

语义信息。因此作者在主干网络之外引入了注意力特征（attentive feature network）分支网络， 该分支网络共包含与主干网络相同的三个分支网络，其中每个分支网络在不同的位置应用多向注意力（multi-direction attention）模块。 应用于不同位置的多向注意力模块能够从不同尺度的特征图中提取有效的图片特征。注意力模块则由两个部件组成，分别为注意力生成模块以及注意力应用模块。注意力生成模块以分支网络特定位置的特征为输入，利用1×1卷积网络压缩特征通道数，生成注意力特征图。注意力应用模块则将注意力特征图与前向和后向的特征进行点乘，实现多向的注意力应用。众所周知，不同网络位置的特征拥有不同层级的特征表达能力，为了融合多尺度的特征，该方法将主干网络与三条分支网络的特征进行融合，得到图片特征的最终表达。由于该方法将主干网络的参数复制了三次形成三条分支网络，因此模型的参数和计算量大幅增加。并且在训练过程中采用了复杂的多阶段训练法。

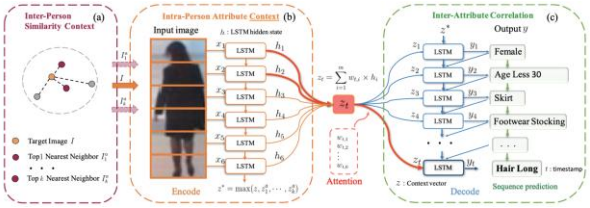

图 10: JRL 方法的模型结构图

为了建模行人属性之间的依赖关系并利用属性的上下文信息，Wang 等人将行人属性识别任务视为一个时序预测任务，并提出了 JRL [17]，其网络结构如图 10 所示。 行人属性之间具有较强的依赖和因果联系，例如“裙子”属性往往和“女性”属性同时出现在一张图片中。 因此，该方法从这一角度出发，采用循环神经网络 [62]（RNN，recurrent neural network）的编码器-译码器 (encoder-decoder) 结构来建模行人属性识别问题。具体来说，在编码器 LSTM 部分， 该方法采用硬划分的方式， 将行人图片划分为六个水平带状区域，形成从图片顶部到底部的区域序列。将该序列特征作为输入送入长短期记忆网络（Long Short-Term Memory，LSTM） [63]，通过长短期记忆网络建模各个区域之间的依赖关系。同时，考虑到相似的行人拥有相同的属性特征，作者选取与目标行人特征最相似的 K 个行人特征，对 $k+1$ 个特征进行进行逐元素取最大值操作作为输入译码器 LSTM 的特征。该方法从序列预测的角度，重新建模了行人属性识别任务，为之后的方法开辟了新的思路。

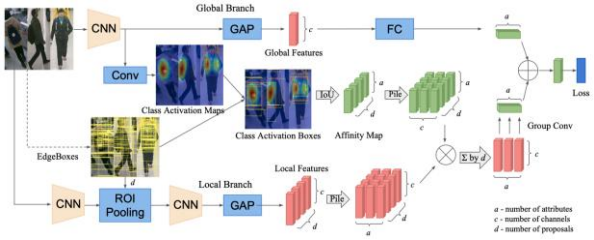

图 11: LGNet 方法的模型结构图

为了避免利用人体视角，姿态，关键点等额外信息，Liu 等人提出了定位引导的属性识别方法 LGNet [50]，其网络结构如图 11 所示。该方法基于预先提取的候选框以及属性位置之间的关系来为局部特征分配特定属性的权重。该方法包含两个分支网络，全局分支网络以及局部分支网络。其中全局分支网络利用输入图片生成所有属性的位置，局部分支利用位置来预测属性的类别。全局分支网络和局部分支网络都是由 Inception-v2 架构改编而来，因为它在多尺度特征提取方面具有很大的扩展性。为了生成可靠的类别激活图，全局分支网络的采用预训练的参数并固定不变。为了提取输入图片的局部特征，作者采用 EdgeBoxes [64] 生成若干个区域候选位置，并利用感兴趣区域池化层提取相应的区域特征。为了衡量不同局部特征对属性类别的贡献程度，作者计算了类别激活图与候选位置之间的交并比 (Intersection over Union)， 并将其作为空间亲和度矩阵。将该亲和度矩阵进行重塑并与全局分支的特征进行融合，构成最终的图片特征，用于属性分类。

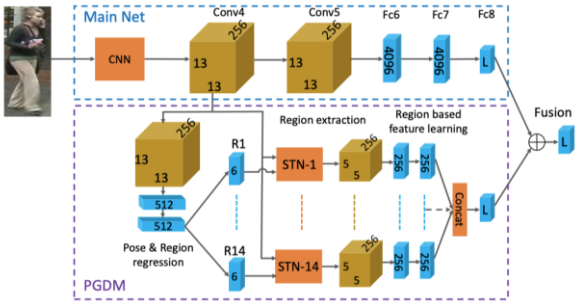

图 12: PGDM 方法的模型结构图

与 LGNet 方法不同，Li 等人认为通过引入人体先验的结构信息能够有效的提高属性识别的性能。因此，作者利用预训练的人体姿态估计模型 CPM(Convolutional Pose Machines) [65] 网络来提取粗粒度的人体结构信息。其网络结构如图 12 所示。Li 等人提出的人体姿态引导的深度模型

PGDM (Pose Guided Deep Model) [51] 包含三个部分，分别是粗粒度的姿态估计，人体部件定位，以及特征融合模块。与之前的方法不同，PGDM 不只是利用预训练模型来提取人体关键点信息，而是将其嵌入到整个模型框架中进行学习。因此，除了主分支作为常规的属性分类网络之外，作者将姿态估计与人体部件定位两个模块引入到额外的分支中来进行人体关键点的回归。其中人体部件定位模块利用空间变换网络 STN （Spatial Transformation Network） 来实现 [66]。STN 为每个关键点预测六个参数，每个关键点相关的区域都采用独立的卷积网络来学习。而特征融合模块则是将学习到的特征进行串接得到最终的图片特征。该方法同样采用了多阶段的训练方式。

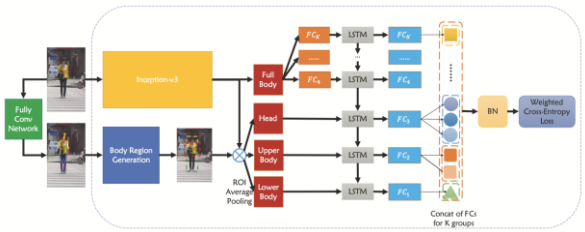

图 13: GRL 方法的模型结构图。

与之前的方法不同，Zhao 等人发现行人属性之间具有明显的空间和语义关系，可以根据空间和语义关系对属性进行分组识别。因此，作者以 RNN 为基础提出了 GRL (Group Recurrent Learning) 方法，该方法的网络结构如图 13 所示。该方法充分利用了组内属性互斥以及组间属性的关系来提高属性识别性能。以 PETA 数据库为例，作者首先将 35 个属性分为 7 组，分别为性别，年龄，头部属性，上半身属性，下半身属性，脚部属性，以及配饰等。为了提取不同属性组对应的局部信息，GRL 方法引入了额外的姿态估计模型 SpindleNet [67] 来获取人体关键点信息。 作者根据人体关键点信息，将人体分为头部特征，上半身特征， 以及下半身特征。 除此之外，作者将全局特征作为其余组别属性的对应特征。在获得每组属性的对应特征之后，作者将特征序列送入 LSTM 网络中，对每次 LSTM 的输出特征，只用其来分类对应组别的属性。另外， 为了缓解属性的样本不均衡现象， 作者在最后输出的预测向量之后添加批归一化层（Batch Normalization, BN） 来进一步提高模型的性能。

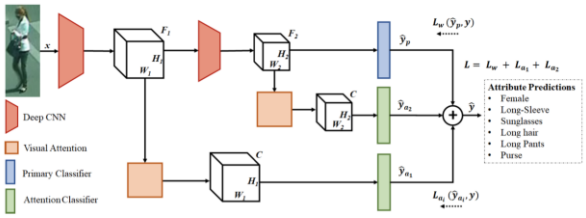

图 14: MsVAA 方法的模型结构图。

考虑到不同的属性样本拥有不同的分布并且人体属性天然存在样本不均衡的现象，Sarafianos 等人 [36] 针对这一问题提出了 MsVAA 方法。该方法的网络结构如图 14 所示。该方法利用了残差网络中各个残差模块输出的多尺度的特征图，通过利用多尺度的空间信息来提取不同粒度的属性特征。 同时， 受类别激活图的影响， MsVAA 在每个阶段引入了注意力机制， 通过模型的梯度回传策略直接学习各个属性的空间注意力图。并对多个空间注意力图进行全局平均池化得到对于属性类别的预测概率。除了模型设计之外，作者将 Focal loss [68] 与 常规的基于正样本概率加权的二元交叉熵损失函数进行结合，通过增加正样本的损失并同时减少负样本的损失来改善样本分布不均衡的问题。

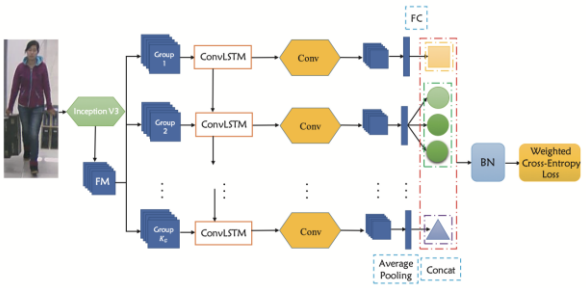

图 15: RC&RA 方法的模型结构图。

受到 GRL 方法将属性进行分组识别思路的影响， Zhao 等人提出了循环卷积（Recurrent Convolution）和循环注意力（Recurrent Attention）模型。 该方法的网络结构如图 15 所示。其中循环卷积模型通过卷积长短期记忆网络 （ConvLSTM） [69] 单元来挖掘不同属性组之间的关系， 而循环注意力模型利用属性组内的空间局部性以及组间的注意力关系来提高行人属性识别的性能。循环注意力模型将循环学习和注意力模型相结合来突出特征图中的空间位置，并挖掘不同属性组之间的注意力关系来获得更精确的注意力区域。在实现方面，作者将属性根据空间位置和语义关系分成若干组，并由此构建出多条分支网络。输入图片首先经

过主干网络，从而提取行人图片的全局特征，该特征则同时送入多条分支网络。其中每条分支网络中的 ConvLSTM 层是时序相关的，从而建模各个属性组之间的关系。同时，对于特定的分支网络，ConvLSTM 层输出的图片特征经过卷积和 Sigmoid 函数的处理生成注意力图。该注意力图作为权重参数， 与原始的图片特征形成残差结构，最后得到的特征则作为该属性组的特征应用于组内属性的分类。 该方法从时序建模和空间建模两个维度出发，分别采用长短期记忆网络以及注意力机制来实现。在该方法中，不同属性组特征送入 ConvLSTM 的次序不同，会导致属性组之间的关系建模的效果不同。 因此， 作者针对属性组建模次序这一变量进行实验， 得出按照全局属性到局部属性的次序来建模能够实现更好的效果。

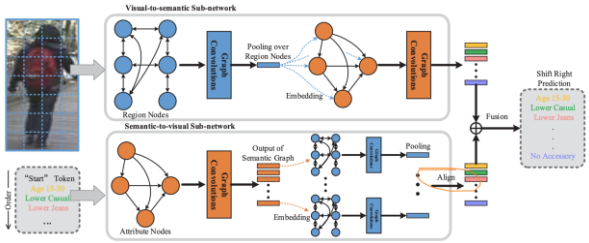

图 16: VSGR 方法的模型结构图。

随着循环卷积网络在行人属性识别领域的应用，越来越多的方法将行人属性识别考虑为一个序列预测问题。Li 等人提出一个视觉语义图卷积的方法 VSGR [54]，其网络结构如图 16 所示。该方法包含一个空间图和语义图， 通过在图上进行图卷积的操作来使空间图能够捕捉区域之间的空间关系， 语义图则可以捕捉属性之间的语义关系。 该方法是图卷积在行人属性识别领域的首次应用。从实现的角度来说，作者构建了一个两分支网络。第一条分支为视觉到语义的子网络，该分支首先将行人图片作为输入，利用主干网络提取特征图，对于特征图中的每个像素点，作者将其作为空间图中的节点，每个节点由空间位置的特征向量所表示。通过在空间图中进行图卷积操作，建模各个空间位置之间的相似度，并利用残差结构以及池化操作得到图片的最终特征表达。第二条分支为语义到视觉的子网络，该分支利用词向量技术 word2vector 得到各个属性对应的语义嵌入向量。通过将每个属性的语义嵌入向量作为语义图中的一个节点从而构建语义图，与空间图上的操作类似，在语义图中进行图卷积等操作从而得到语义表征的特征向量。最后将两条分支的特征向量进行融合，进行属性分类。

该方法不仅仅将行人图片作为输入，同时将属性的语义词向量作为输入，增加了输入信息，实现更好的属性识别性能。

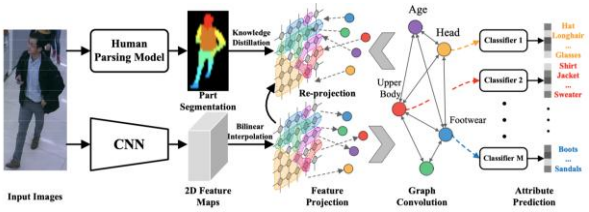

图 17: GRKD 方法的模型结构图。

继 VSGR [54]之后，Li 等人又提出 GRKD [55] 方法，如图 17 所示。 该方法延续了之前 VSGR 从语义和空间两个角度建模属性识别问题的思路，并在此基础上引入了知识蒸馏（Knowledge Distillation）[70] 和 人体部件语义分割的信息。 其中作者 利用 SPReID [71] 分割模型得到输入图片的语义分割结果。 将主干网络得到的图片特征进行图卷积操作之后， 映射到语义空间节点中。 最后利用每个节点的特征对相应的属性组进行分类预测。

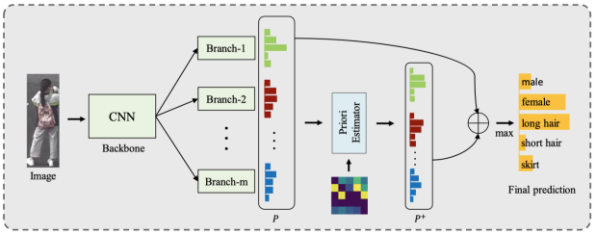

图 18: AAP 方法的模型结构图。

Han [56] 等人充分利用了属性之间的相互关系，并提出了 AAP 这个多分支网络， 其网络结构如图 18 所示。 受到卷积网络在底层中学到的是通用的视觉表观信息，而在高层中能够学到语义信息这一特性的启发， 该结构采用底层参数共享，高层参数独立的设计思路。 具体地，该方法将主干网络得到的图片特征分为四部分，分别是全局特征，头部特征，上半身特征，以及下半身特征。其中特征的划分采用人为硬划分的方式得到。与之前对属性进行分组识别的方法不同，该方法在不同的分支中，利用不同位置的空间特征，对所有属性类别进行分类预测，最后对各分支分类预测的结果取最大值。除此之外，作者还从概率的角度分析了数据库中各个属性之间的共现先验概率（co-occurrence prior probability ）。并将其与模型的学习过程将结合，对属性预测概率进行约束，使得模型的预测结果更符合数据库的先验分布。

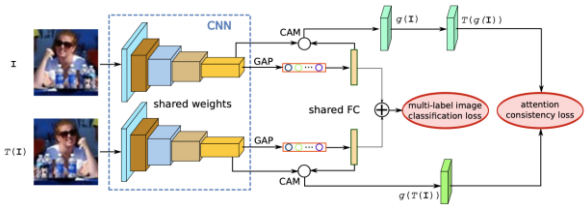

图 19: VAC 方法的模型结构图。

除了考虑属性的先验约束，图片之间的先验约束也被最近的方法纳入到模型的训练过程中。Guo 等人依据同一张图片不同数据增广样本之间，模型的注意力区域应该保持一致的假设，提出了视觉注意力一致性方法 VAC [27]，其网络结构如图 19 所示。该方法将人类的先验转换为注意力区域的约束条件，并将其与分类损失相结合作为模型的监督信号。 以水平随机翻转的数据增广方式为例，原始图片的注意力区域应该与增广图片的注意力区域水平对称。 因此，作者提出了一个两分支网络，两条分支共享网络参数。其中一条分支用来对原始图片进行特征提取，并得到网络在原始图像上的注意力区域。另一条分支用于对增广图片进行特征提取，并得到网络在增广样本上的注意力区域。在损失计算方面，除了常规的二元交叉熵损失之外，作者根据原始样本和增广样本注意力区域之间的距离，提出了注意力一致性损失，约束模型注意到更具判别性的属性区域。除了水平随机翻转之外，作者还对缩放变换，平移变换，以及旋转变换等增广方式进行了分析和实验。

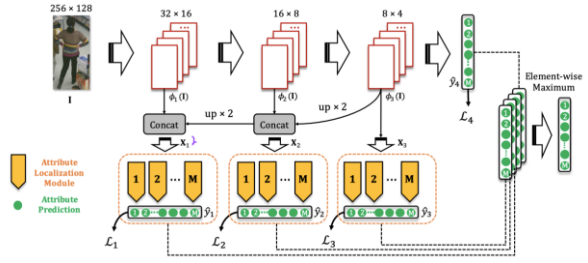

图 20: ALM 方法的模型结构图。

由于多尺度特征在目标检测，分割等领域的广泛应用，不少工作也将多尺度特征应用在行人属性识别方法中。 Tang 等人提出的 ALM [19] 方法，通过利用主干网络各级不同尺度的特征图以及特征金字塔网络（Feature Pyramid Network， FPN）[72], 并结合空间变换网络 STN 以及 压缩激励模块（Squeeze and Excitation block，SE block）来提取每个属性对应的特征并进行单独分类。网络结构如图 20 所示。ALM 网络以残差网络 Resnet50 作为主干网络，并通过特征金字塔网络融合不同阶段不同尺度的特征图。对于不同尺度的特征图，首先对分辨率较低含有较丰富语义特征的特征图进行双线性插值上采样，并与分辨率较高含有视觉特征较丰富的特征图进行合并，构成最终的特征图。在此基础上，作者利用 SE 模块，对特征图中各个通道的特征进行加权融合并通过 STN 网络为每一个属性生成单独的特征向量。因此，该网络共计在三个尺度的特征图上，分别对每个属性进行预测。另外，作者也通过主干网络对属性进行概率预测。将四部分的概率预测值进行逐元素取最大值操作得到最终的分类结果。

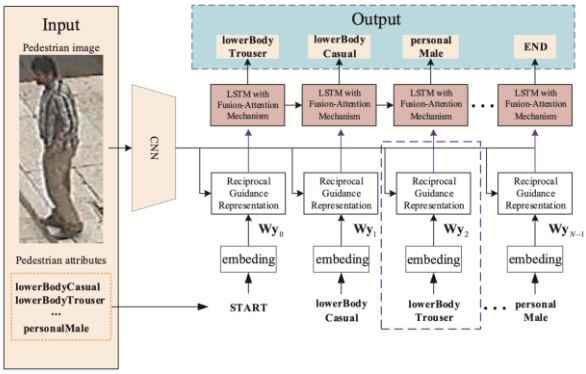

图 21: IA^2-Net 方法的模型结构图。

Ji 等人提出的图片属性互相指导的注意力网络 IA^2-Net [57]，其网络结构如图 21 所示。 该网络利用图片引导的特征和属性引导的特征来对不同的属性学习不同综合特征。为了自适应的优化特征的权重分布，作者又提出了一个混合注意力机制。该模型将图片和属性的 one-hot 特征向量作为输入，并从 one-hot 特征向量中生成两个不同的隐空间。其中一个隐空间特征和输入特征一起，生成图片引导的特征。另一个隐空间特征和输入的 one-hot 特征一起，生成属性引导的特征。图片引导的特征和属性引导的特征进行拼接送入 LSTM [63] 网络，来建模各个属性之间的关系。在损失层面，作者使用分类损失和 Focal loss 损失 [68]。

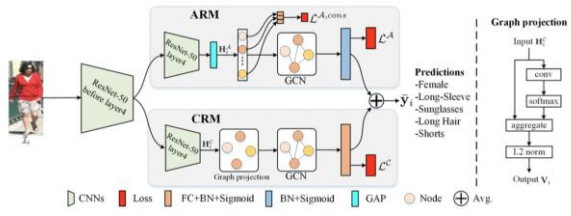

图 22: JLAC 方法的模型结构图。

受到多分支网络和图卷积网络的启发，Tan 等

人提出了属性和上下文关系联合学习的框架 JLAC [34] 网络， 该网络结构如图 22 所示。该网络共包含属性关系模型和上下文关系模型。对于属性关系模型，作者用属性特征为节点构建了属性图，并通过受约束的损失以及图卷积操作来学习属性之间的关系。对于上下文关系模型，作者首次提出一个图映射方法来将 2D 特征图中不同的空间区域映射为一系列的图节点，并结合图卷积网络来建模这些区域的上下文关系。具体地，该方法采用 Resnet50 作为主干网络，其中 layer1 到 layer3 层作为共享层，两个 layer4 层分别用来提取属性关系模型和上下文关系模型的初试特征。在属性关系分支中，对于初试特征，作者利用多个全连接层得到每个属性特征，这些属性特征一方面直接送入分类器得到属性预测结果，另一方面作为节点构建属性图网络，通过图上的卷积操作进行属性特征的更新，最后再送入分类器。在上下文关系分支中，作者借鉴 NetVLAD [73] 方法，将特征图中像素点的特征进行聚类融合， 作为若干个节点构成图网络， 通过图卷积操作来更新节点特征，最后将所有节点特征进行拼接送入分类器。 作为目前性能最高的方法， 该方法在两条支路上利用了图卷积操作建模不同特征之间的关系。

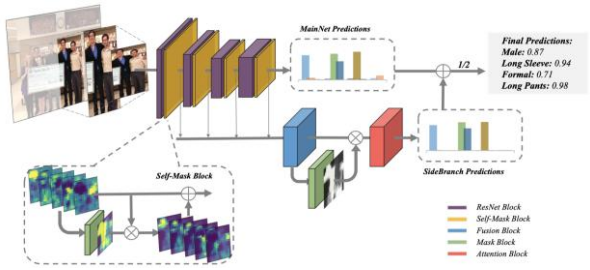

图 23: Da-HAR 方法的模型结构图

为了进一步提高模型对属性相关区域的定位能力， Wu 等人提出了多分支的 Da-HAR [23] 网络， 如图 23 所示。通过逐级的注意力机制来学习有判别性的对属性特征。该网络的学习过程分为两步，首先在网络的输入阶段引入自遮挡（self-mask）模块，该模块通过堆叠的1×1卷积层来预测图片的显著性信息， 初步过滤掉背景等无关信息， 从而实现初步的注意力机制。经过自遮挡模块的图片作为输入，通过主干网络输出每个属性的预测概率。同时，分支网络，将主干网络各个尺度的特征图进行融合，并送入分支网络的遮挡注意力（masked attention）模块。该模块借鉴了 SRN 网络中的置信度加权注意力机制，在自遮挡模块提取的粗粒度特征的基础上进一步提取精细化的属性特征。另外，该模型利用在 COCO [28] 数据库上训练得到的分割模型 FPN [72]， 得到每张输入图片的分割结果作为监督信号来训练遮挡注意力模块。因此该模型通过在分支网络中引入分割任务来形成一个多任务学习框架，来提高模型对属性相关区域的定位能力。

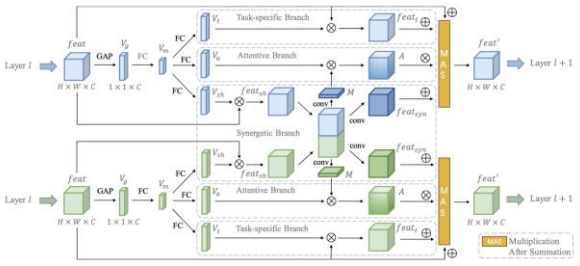

图 24: CAS 方法的模型结构图。

之前的方法往往集中于从空间角度对特征向量进行利用，而忽略了特征向量通道角度的利用。Zeng 等人提出的 CAS [58] 网络从多任务学习的角度提取了有判别性的特征通道和空间区域来学习更有效的特征表达。如图 24 所示，CAS 网络包含两个参数不共享的 Resnet34 网络，并在两个网络的 layer1 到 layer4 层中加入注意力共享（attention sharing）模块。 该注意力共享模块的输入为两个网络的特征，每个模块分别包含三条支路，其中在特定任务支路（specific-task branch）和注意力支路（attentive branch）中，两个模块彼此独立，而在协同支路（synergetic branch ）中两个模块互相协作。 两个 ResNet34 网络层的输出经过全局池化以及全连接层得到通道注意力向量，并作为输入送入到各自注意力共享模块的三条支路。 其中两条协同支路分别将通道注意力向量对原始特征进行加权融合， 并将得到的特征进行拼接。该过程融合了两个 ResNet34 网络的不同特征表达。对拼接后的特征进行降维操作生成空间注意力图并分别应用于各自的注意力支路。对于特定任务支路，则将通道注意力向量对原始特征直接加权。最后将三条支路得到的特征进行相加，送入下一阶段网络。 尽管该网络使用了双倍的计算量和参数量，但并没有取得较高的性能。

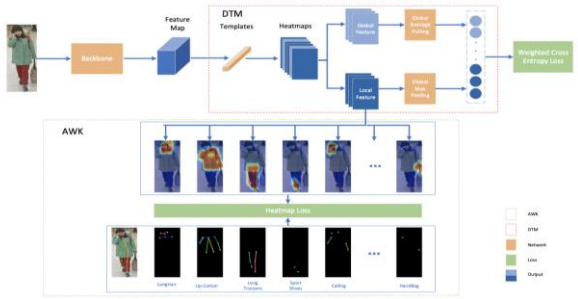

图 25: DTM 方法的模型结构图。

Zhang 等人提出的 DTM [59] 通过精确的捕捉人体部件特征来提高模型的识别性能。如图 25 所示，该方法在训练阶段主要分为主干网络以及人体关键点监督的辅助网络。在主干网络中，作者利用可学习的 $C\times N$ 卷积核， 其中 N 代表属性的数量， 从主干网络的输出特征图中提取每个属性的热力图（heatmap）， 其中每个热力图中的值反应的是该属性出现在对应位置的概率。 得到的热力图被送入两条分支网络， 其中全局特征分支网络利用全局平均池化得到对应的每个属性的概率预测，用于预测全局属性，例如年龄，性别，体态等； 局部特征分支网络利用全局最大池化得到对应的每个属性的概率预测，用于预测局部属性，例如发型，服饰，配饰等。 两条支路的概率预测在最后进行合并作为最后的预测结果。 除了利用属性标签作为监督信号之外， 作者还利用人体关键点估计模型 [74] 生成的十四个关键点信息作为辅助监督，用于监督局部特征分支中的热力图学习。 因此，该方法在之前方法采用多分支模型的基础上， 利用人体关键点，强化对属性相关区域的特征提取。

Jia 等人提出的对于行人属性识别的重思考工作[15] 包括对现有数据库的分析以及现有方法的分析两部分。 在对现有行人属性数据库的分析方面，作者认为目前主流的三个数据库其中的两个，由于行人图片采集方式的局限性，存在数据泄露的问题，使得测试集不能充分的衡量模型的泛化性能，并且普遍高估了目前主流的各种方法。现有的行人属性识别方法，由于其基准方法性能存在较大差异，使得各个方法之间的比较非常困难，不能直观的量化各种方法的创新性。针对这两个问题，作者在数据库方面，基于 PETA 和 RAP 两个数据库，根据行人身份重新划分了训练集和测试集， 使得训练集和测试集的行人身份符合零样本的设定。在方法方面，作者复现了具有代表性的算法， 并在传统数据库和新提出的数据库上进行测试验证。实验表明， 作者新提出的数据库能够有效衡量各种算法的泛化性，并且能够对不同的算法进行直接的比较。 除此之外，作者还提出了一个强有力的基准方法， 该方法只使用一个主干网络就能实现和现有最先进的方法接近的性能， 为之后的行人属性方法提供了一个参考。

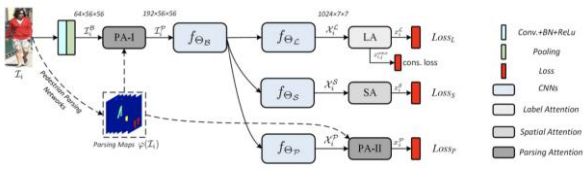

图 26: JLPLS-PAA 方法的模型结构图

为了综合考虑各种注意力机制在行人属性识别领域的应用，Tan [34] 等人提出了 JLPLS-PAA 方法。 如图 26 所示，该方法主要包括三部分， 分别为人体部件分割注意力网络，标签注意力网络，以及空间注意力网路。这三个不同的网络分别从不同的角度来捕捉有效的行人特征。 人体部件分割注意力网络不仅仅通过学习将注意力聚焦在哪里，也通过学习如何融合不同身体部件语义区域中特征来提取有判别性的特征。 标签注意力网络的目的是有针对性从每个属性中收集判别信息。 与人体部件分割网络和标签注意力网络不同， 空间注意力网络从全局的角度考虑问题，为每个属性选择若干个重要的具有判别性的图片区域以及像素点。作者将三个网络有机的结合在一起形成了类似多任务的联合学习框架，使其能够并行的提取互补和互相关联的特征。 为了能够对人体部件分割网络进行有效的训练， 作者对行人再识别数据库 VIPeR [75] 进行了标注， 提供其属性标签和语义分割结果。 同时，作者采用了性能更强的主干网络 SE-BN-Incep 网络以实现更高的性能。

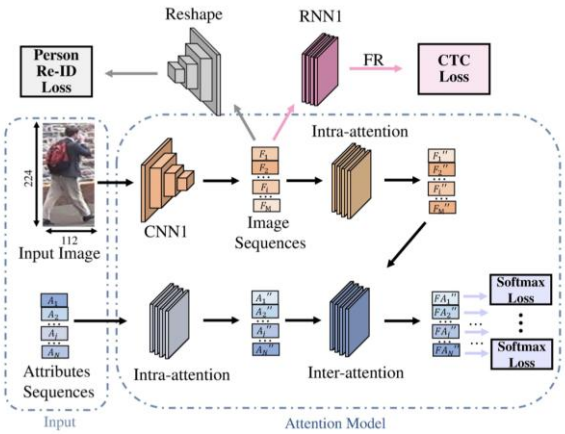

图 27: SCRL 方法的模型结构图。

Wu [35] 等人提出的 序列上下文关系学习模型 SCRL 分别对空间-语义关系，空间上下文，以及语义联系等三种关系进行建模。 如图 27 所示，该方法采用双分支网络， 每条分支的输入分别为图片信息和属性信息。在图片分支中，首先将输入图片映射为特征向量序列，并以序列之间的相似度为依据，构建自注意力模块来更新特征序列，为了融合特征序列中不同特征包含的图片信息，作者采用 RNN 结构，以特征序列为输入，将 RNN 的输出作为最终的图片特征， 并以自然语言处理中的 CTC 为损失函数对模型的训练过程进行监督。 在属性分支中， 将所有属性的特征向量进行拼接，并通过全连接层得到不同属性的特征表达， 通过自注意力模块和与图片特征的相互注意力模块将属性特征和图片特征进行加权融合，并采用常规的交叉熵损失进行监督。除此之外，作者还引入了行人身份信息的监督。

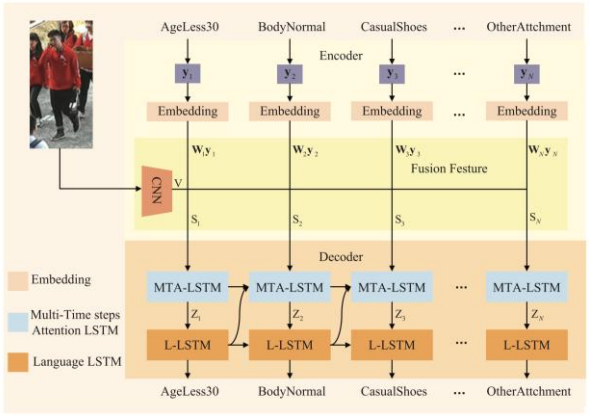

图 28: MTANet 方法的模型结构图

为了解决图片与属性之间的复杂联系以及属性之间的不平衡分布问题，Ji 等人提出了 MTANet 方法 [60]。 如图 28 所示，该方法提出了一个新的多步注意力网络来强化关系的建模。与之前的方法只关注循环神经网络中当前时间阶段与之前时间阶段不同，该方法利用了循环神经网络中下一时间阶段中的信息。通过自适应的捕捉多个时间阶段里循环神经网络中的特征，能够利用更多上下文信息。同时。为了缓解行人属性不均衡分布的影响，一个基于 Focal loss [68]的类别平衡的损失函数被提出。

表 4 监控场景中的行人属性识别方法在 PETA, RAPv1, PA100K 三个数据库上的性能。其中 “--”表示原文未展示相关结果

方法	额外信息	框架结构	主干网络	特点部件	目标问题
DeepMAR[11] (ACPR2015)	无	单分支单尺度	CaffeNet	–	首次引入深度学习；并行训练多个属性
WPAL[48] (BMVC2017)	无	单分支多尺度	GoogleNet	SPP[61]	利用弱监督目标检测中的定位方法来辅助属性定位
VeSPA[49] (BMVC2017)	无	多分支单尺度	GoogleNet	–	探索利用行人视角信息来辅助行人属性识别
HPNet[14] (ICCV2017)	无	多分支多尺度	InceptionNet	–	多尺度多级别的注意力信息的融合
JRL[17] (ICCV2017)	否	单分支单尺度	AlexNet	LSTM[63]	建模属性语义特征之间的关系
LGNet[50] (BMVC2018)	否	多分支单尺度	Inception-v2	EdgeBox[64], CAM[76]	属性相关区域特征的挖掘以及重加权
PGDM[51] (ICME2018)	人体关键点估计模型	多分支单尺度	CaffeNet	STN[66]	利用人体结构先验定位属性相关区域以及区域特征的融合
GRL[52] (IJCAI2018)	人体关键点估计模型	多分支单尺度	Inception-v3	RoI pooling[77] + LSTM[63]	利用人体结构先验定位属性相关区域
MsVAA[36] (ECCV2018)	无	多尺度多分支	ResNet101	Focal loss[68]	利用多尺度特征的有效性；解决样本不均衡问题
RA[53] (AAAI2019)	无	单尺度多分支	Inception-v3	ConvLSTM[69]	属性组关系建模

VSGR[54] (AAAI2019)	无	单尺度多分支	ResNet50	GCN[78]	属性空间特征以及语义特征的关系建模
VRKD[55] (IJCAI2019)	人体部件分割模型	单尺度单分支	ResNet50	GCN[78]	模型对属性的空间定位能力；属性语义特征关系建模
AAP[56] (IJCAI2019)	无	单尺度多分支	ResNet50	–	区域特征的有效提取；利用先验概率约束属性的分布
VAC[27] (CVPR2019)	无	单尺度多分支	ResNet50	数据增广方式	空间注意力的一致性
ALM[19] (ICCV2019)	无	多尺度多分支	BN-Inception	STN[66], FPN[72], SE block[79]	在各个尺度的特征图上进行属性相关区域特征的提取
IA2*Net*[57] (PRL2019)	无	单尺度单分支	ResNet50	LSTM[63]	利用视觉外观和属性之间的关系来学习综合特征
JLAC[20] (AAAI2020)	无	单尺度多分支	BN-Inception	GCN[78], NetVLAD[73]	建模属性关系；建模上下文关系；
Da-HAR[23] (AAAI2020)	目标分割模型	单尺度多分支	ResNet50	–	模型对属性相关区域的定位能力
CAS[58] (ICME2020)	无	单尺度单分支	ResNet34	–	利用特征向量中通道维度的信息
DTM[59] (Arxiv2020)	人体关键点估计模型	单尺度单分支	ResNet50	–	人体部件特征的精确捕捉
JLPLS-PAA[34](TIP2019)	人体部件分割模型	单尺度多分支	SE-BN-InceptionNet	–	不同维度的注意力网络的结合
SCRL[35] (TCSVT2020)	行人再识别模型	单分支单尺度	ResNet50	RNN[80]	空间特征-语义特征之间，空间特征之间，语义特征之间的关系建模
MTANet[60] (PRL)	无	单分支单尺度	ResNet152	LSTM[63]	属性之间的关系建模；属性之间的不平衡分

为了进一步对现有的监控场景中的行人属性识别方法进行分类，我们从现有方法是否利用额外信息，框架结构，主干网络，特点部件，以及方法所解决的目标问题等五个方面对现有方法进行梳理，见表 4。是否利用额外信息是指在模型训练过程中，是否涉及除行人图片和属性标注的其他先验知识，例如人体关键点 （human key points），人体部件分割结果（human parsing），或者行人身份标注等。现有方法主要从两个方面利用额外信息，分别是利用额外信息增强模型对属性相关区域的定位[51, 52, 55]以及利用额外信息提供监督信号,引入新的损失函数[34, 35]。框架结构是指方法提出的模型整体框架结构，主要分为单分支单尺度 ，单分支多尺度，多分支单尺度，多分支多尺度四类。其中分支（branch）是指模型并行与主干网络的支路，尺度（scale）是指所利用的特征图的大小。在相同主干网络的情况下，采用单分支单尺度框架结构的方法 [11, 15, 21, 35] 较为简洁，复杂度低计算量少，易于部署和实际应用，但是没有充分利用不同尺度特征图中包含的不同级别的语义信息。而采用多分支多尺度框架结构的方法 [19, 34, 36]，复杂度高计算量大，能够充分探索和利用各个级别的特征图中丰富的语义信息。主干网络（backbone network）是指模型采用的用于提取图片特征的网络，目前主流的主干网络有 InceptionNet [81, 82] 和 ResNet [83] 两个系列。特点部件是指该方法中采用的有代表性的网络模块，特点部件可以有效的识别方法的核心贡献。目标问题是指该方法所尝试解决的监控场景中行人属性识别技术存在的问题。目标问题大体可以分为三类，第一类是解决模型对属性相关区域定位不准确的问题 [14, 19, 23, 27, 50-53, 55, 59]，该类方法尝试利用额外信息辅助或者注意力机制等来提取属性的判别性特征；第二类是解决属性关系建模的问题 [17, 20, 35, 52, 54]，该类方法的特点是将监控场景中的行人属性识别这一分类任务转化为序列预测任务，利用 RNN, LSTM，GCN 等模块

来实现；第三类是解决属性之间存在的样本分布不平衡问题 [12, 20, 36]，这类方法的特点是在原有的分类损失的基础上引入各个属性的权重。

通过对上述方法的分类与总结可以发现，目前监控场景中的行人属性识别领域依然存在模型对于属性相关区域定位能力较弱的问题，这主要是由于作为分类任务，监控场景中的行人属性识别任务缺少位置监督，只有类别监督，而模型对于分类任务更侧重于定位判别性的区域而不是属性存在的整体区域。另外，目前的属性数据库图片数量较少，噪声标注较多，这些因素都导致了模型容易出现过拟合现象，不利于模型精确的定位属性的相关区域。另外，如何有效的建模属性之间的语义关系依然是一个尚未解决的开放问题，尽管一些方法利用序列预测网络来解决属性之间的依赖关系，但是其对于性能的提升依然非常有限，同时引入了较高的计算量和复杂度。除此之外，由于部分数据库中存在数据泄露的问题 （如第三章所述），现有方法的性能不能准确衡量方法的泛化性能，因此通过目前数据库上的性能来对各个方法进行比较也存在局限性。同时，现有的方法都局限在同一数据域上进行训练和测试，没有考虑到监控场景中的行人属性识别任务的实际需要，缺少跨域设定下的对比。

4.2 人体属性识别方法

本节将以在 WIDER 数据库上经过评测的方法为主线， 梳理人体属性识别任务的发展过程。方法在 WIDER 数据库中各个属性上的性能如表 5 所示。其中 VeSPA, MsVAA，Da-HAR 等方法既可以用于人体属性识别也可以用在监控场景中的行人属性识别任务上， 因此在本节中不再赘述。

Fast RCNN [77] 作为一个通用物体检测器被 Girshick 等人提出，该方法首先通过卷积层提取输入图片的特征图，并利用感兴趣区域池化层（RoI pooing layer）从特征图中提取每个候选物体（object proposal）的特征向量。不同候选物体拥有相同维度的特征向量。最后，将特征向量送入分类层与位置回归层来得到物体的最终类别和位置。R*CNN [84] 方法利用 Fast RCNN 网络通过提取上下文信息来解决行为识别任务。WIDER 数据库的作者将 Fast RCNN 和 R*CNN 两种方法在 WIDER 数据库上进行微调（fine-tuning）得到其在属性识别任务上的性能。

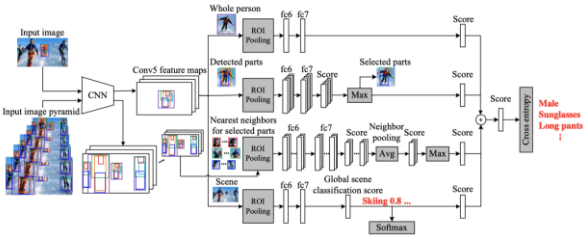

图 29: DHC 方法的模型结构图

Li 等人提出 DHC [22]方法，如图 29 所示，该方法以 Fast RCNN 为主干网络框架， 通过利用高斯金字塔生成多尺度的输入图片，从而得到多尺度的特征图。在主干网络之后，作者提出四条分支网络分别得到不同级别的特征信息。首先，第一条分支网络利用 Fast RCNN 得到的人体检测框，并经过感兴趣区域池化层得到完整的人体特征。第二条分支网络利用身体部件检测网络 Poselet [44] 得到人体五个部件区域的检测框，并经过感兴趣区域池化层得到五个人体部件的局部特征。第三条分支网络则在多尺度特征图中获取距离目标人体各个部件最近的其他人体的部件，通过对这些部件特征的加权平均得到以人为中心的上下文信息特征。第四条分支网络则将全局特征作为输入，通过对场景的分类，得到背景信息的上下文特征。其中，前三条分支网络获取的特征通过加权融合，来进行人体属性的分类。因此，该方法利用人体及部件检测器，并结合多尺度的输入特征以及场景类别，实现属性分类性能的提升。

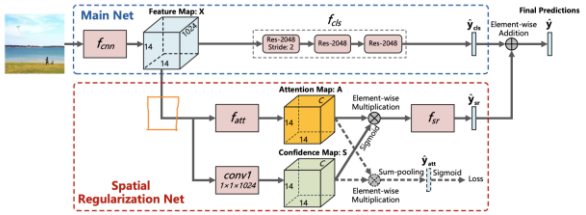

图 30: SRN 方法的模型结构图。

为了解决之前的方法过于依赖人体部件检测网络性能的问题，Guo 等人提出了基于类别激活图（Class Activation Map， CAM）[76]的全局注意力属性识别的方法[85]。 如图 30 所示，该方法利用类别激活图定位特征图中的属性相关区域， 并通过设计新的损失函数使类别激活图覆盖的区域更小更集中来实现更精确的属性定位能力。对给定输入样本的特定属性，该方法首先将对应的类别激活图经过 Softmax 归一化处理， 从而得到类别概率图， 再对类别概率图经过全局最大池化（global average pooling ）得到该属性的最大概率值。 新的

损失函数通过强化该概率值来实现模型更精细的更集中的属性定位能力。在训练过程中，首先采用常规的交叉熵损失进行监督训练，再用新的损失函数微调模型的参数。该方法避免了引入额外的人体部件检测模型的开销，在不增加模型参数和计算量的前提下，实现了属性识别性能的提高。

人体属性识别任务做为一个多标签分类任务，其图片中存在的各个人体属性之间存在较强的依赖关系。Zhu 等人通过提出 SRN [86] 网络来建模各个属性之间的语义和空间关系。SRN 网络将主干网络 ResNet101 输出的特征图作为输入，利用注意力机制（attention mechanism）来提取各个属性之间的空间和语义关系。其中空间关系和语义关系分别通过引入新的可学习的卷积网络来学习各个属性的空间位置和语义特征。该方法采用四阶段训练方式得到最终的模型。该方法的优势在于不需要引入额外的监督信息就可以学习各个属性的空间和语义关系。 但是缺点在于模型引入了额外的参数和计算量， 并且采用了复杂的训练过程和数据增广方式。

表 5 人体属性识别方法在 WIDER 数据库上的性能结果。其中“-”表示原文未展示相关结果。

方法	主干网络	mAP
R-CNN (*ICCV'15*)[77]	VGG16	80.0
R*CNN (*ICCV'15*)[84]	VGG16	80.5
DHC (*ECCV'16*)[22]	VGG16	81.3
VeSPA (*BMVC'17*)[49]	GoogleNet	82.4
CAM (*PRL'17*)[85]	VGG16	82.9
ResNet101(*CVPR'16*)[83]	ResNet101	83.7
SRN* (*CVPR'17*)[86]	ResNet101	85.1
MsVAA (*ECCV'18*)[36]	ResNet101	86.4
Da-HAR (*AAAI'20*)[23]	ResNet101	87.3

通过对上述方法的简要概述，目前人体属性识别方法从解决问题的角度，可以分为两大类。 以 DHC [22] 和 Da-HAR [23] 为代表的方法通过解决模型对属性区域定位不准的问题来提高属性识别的性能；以 MsVAA [36] 为代表的方法致力于解决人体属性中样本分布不均衡的问题，来提高模型对少样本属性的泛化能力。从是否引入额外监督信息的角度，可以分为两大类。以 SRN [86] 和 MsVAA [36] 为代表的方法借鉴了类别激活图，使用自注意力机制来提取模型对属性的定位区域；而以 DHC [22] 和 Da-HAR[23] 为代表的方法， 分别使用人体部件检测网络和通用物体分割网络来提供额外的监督信息，从而辅助模型对属性相关区域的定位。

4.3 面向行人再识别的行人属性识别方法

本节对面向行人再识别的行人属性识别方法进行简要概述，其在相关数据库上的结果如表 6 所示。

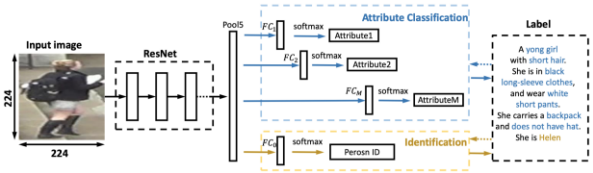

图 31: APR 方法的模型结构图

为了辅助行人再识别任务，面向行人再识别的行人属性识别任务由 Lin [3] 等人首次提出。 作者认为行人属性识别任务和行人再识别任务在特征粒度方法存在差异，行人再识别任务注重行人图像的全局特征，而行人属性识别任务则注重行人图像的局部特征。 之前的行人再识别方法通过构建行人正负样本对来增加同一行人图片间的相似度，同时减少不同行人图片间的相似度。因此，与之前的方法不同，作者从另一个视角思考了如何利用属性类别标签来辅助行人再识别任务。因此，作者基于 Market1501 [39]和 DukeMTMC [40] 两个行人再识别数据库，对其增加标注了属性标签。在方法层面，作者引入了多任务分类模型 APR 网络，如图 31 所示，该方法从同一个全局特征出发，分别得到行人身份特征和不同的行人属性特征，并以行人身份和行人属性标签为监督进行训练。

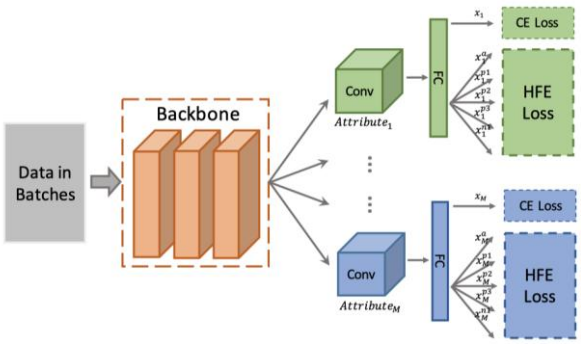

图 32: HFE 方法的模型结构图。

继 Lin [3] 等人提出面向行人再识别的行人属性识别任务之后，Yang [87] 等人提出了级联特征嵌入的 HFE 网络。如图 32 所示，该方法在利用不同分支网络提取不同属性特征的基础上，充分研究了两组特征之间的距离关系，并基于两组特征的距

离关系和三元组损失的基础上，提出了两个新的度量损失。其中基于不同行人同一属性特征之间的距离以及不同行人不同属性特征之间的距离，提出了类间损失函数。基于同一行人同一属性特征之间的距离以及不同行人同一属性之间的距离，提出了类内损失函数。除了相对距离之外，作者还对不同行人不同属性特征之间的距离做了直接的约束。结合原有的分类损失函数，作者以四个损失函数为监督进行网络的训练。

表 6 面向行人再识别的的行人属性识别方法在 Market1501 和 DukeMTMC 数据库上的性能。

方法	主干网络	Market1501					DukeMTMC				
		mA	Accu	Prec	Recall	F1	mA	Accu	Prec	Recall	F1
DeepMAR[11](*ACPR2015*)	CaffeNet	88.68	69.65	82.60	80.24	81.40	86.90	70.67	81.82	82.24	82.03
HPNet[14] (*ICCV2017*)	InceptionNet	91.13	74.82	85.26	83.31	84.27	91.13	67.63	82.77	75.19	79.79
MsVAA[36] (*ECCV2018*)	ResNet101	91.27	75.03	85.64	83.18	84.39	91.27	74.02	84.85	83.44	83.14
APR[3] (*PR2019*)	ResNet50	88.93	70.25	83.52	78.96	81.18	87.88	70.10	82.74	79.02	80.83
HFE[87] (*CVPR2020*)	ResNet50	92.90	78.01	87.41	85.65	86.52	92.90	76.68	86.37	84.40	85.37

4.4 行人属性识别及其相关任务的方法总结与对比

监控场景中的行人属性识别方法，面向行人再识别的行人属性识别方法以及人体属性识别方法，由于其各自针对的任务不同，在方法的总体设计上也有所差别。监控场景中的行人属性识别方法从单分支单尺度网络结构发展到多分支多尺度网络结构，探索利用丰富的属性空间特征并同时建模语义特征之间的联系，结合图卷积网络以及循环卷积网络等方式，提升属性特征的表达能力；对于面向行人再识别的行人属性识别任务，目前的研究方法比较少，主要是因为该任务主要用于辅助提高行人再识别的性能。对于人体属性识别方法，大多采用类别激活图以及人体部件分割，人体关键点估计等模型加强分类模型对属性相关区域的定位能力，提取有判别性的属性空间特征，目前的方法并没有充分利用属性特征之间的关联。

5 行人属性识别的总结与展望

5.1 总结

综上所述，本文对行人属性识别的研究现状进行了总结，一是给出了监控场景中的行人属性识别的任务范畴与定义，并与相似的两个任务进行对比，二是对监控场景中的行人属性识别，人体属性识别，以及面向行人再识别的行人属性识别等三个任务涉及的数据库进行了分析和概括，三是针对行人属性识别问题提出的方法进行了梳理。通过总结和对比可以发现，监控场景中的行人属性识别由于其在图像采集方式和属性标注准则等方面的特点，使得该任务与人体属性识别和面向行人再识别的行人属性识别等任务有明显的不同。但是目前的学术研究中并没有着重区分这三者之间的关联，因此本文从任务产生的背景以及数据库的构成两方面厘清了三个任务之间的区别。通过对数据库和方法的综述，可以发现目前的大多数方法都致力于解决行人属性识别中两大问题。一个是利用注意力机制，额外的人体结构信息，一致性约束等方法，来解决模型对属性相关区域定位能力较差的问题。另一个是采用长短时记忆网络，图卷积网络等方法来解决属性之间关系建模不充分的问题。

5.2 展望

行人属性识别尽管取得了一些成绩，但是诸如属性特征解耦不充分，属性之间的关系建模不完善，以及跨域等问题依然没有得到解决。

首先是属性特征解耦不充分导致不同属性之间的特征相似性过高，缺少判别性。由于网络的输入是一张图片，但网络的输出需要对图片中包含的多个属性进行预测。因此，如何从图片的全局特征中解耦不同属性的特征是未来的一个工作方向。目前该问题的解决方法主要是通过强化模型对图片中各个属性相关区域的定位，例如采用类别激活图，引入部件分割模型或者人体关键点信息等，将人体划分为多个不同的区域，对不同区域中涉及的属性进行分类。但是目前的方法一方面依赖于外部引入的额外信息的准确度，另一方面大多方法只是探索了如何从空间角度对属性特征进行解耦，而没有从特征的通道角度进行探索。因此，如何有效的

从一张图片的全局特征中解耦出有判别性的属性特征是未来的一个工作方向。

其次是属性之间的关系建模不完善。不论是在监控场景中的行人属性识别还是在通用场景中的人体属性识别，属性之间都存在明显关系。但是如何建模这种关系，是从概率的角度建模共生关系，还是从因果图的角度建模属性之间的依赖关系还有待研究。目前的主流方法局限于对属性的空间关系，即像素点特征之间的关系进行建模，加权融合不同空间位置的特征并更新从而实现更有效的属性特征表达。对属性的语义关系的建模则主要是从词向量的角度出发。但是这两种方法对模型的性能提升均不明显。

最后是行人属性识别中的跨域问题。不同场景中的背景，光照，图片分辨率，以及行人的衣着服饰等有着很大的差异。例如，同样属于帽子属性，训练集中可能会出现安全帽，贝雷帽，遮阳帽等，但是在测试集中却大量存在着鸭舌帽，斗笠，以及绒线帽等。因此，同一属性类别的不同实例之间存在显著的外观差异。该问题对于监控场景中的行人属性识别尤其重要，因为监控场景中的行人属性预测一定是针对未见过的行人图片。如何能够使网络具有良好的泛化能力，也是未来的一个研究重点

参 考 文 献

[1] Huang K, Chen X, Kang Y. Intelligent Visual Surveillance: A Review. Chinese Journal of Computers 2015, 38(6):1093-1118. (黄凯奇, 陈晓棠, 康运锋, 等. 智能视频监控技术. 计算机学报, 2015, 38(6):1093-1118.)

[2] Li Dangwei. Pedestrian Attribute Recognition and Re-identification in Surveillance Scenarios[Ph.D. Thesis]. Institude of Automation , Chinese Academy of Sciences. 2018.（李党伟. 监控场景下行人属性识别与再辨识关键问题研究[博士论文]. 北 京: 中国科学院自动化研究所. 中国科学院大学, 2018.）

[3] Lin Y, Zheng L, Zheng Z, et al. Improving person re-identification by attribute and identity learning. Pattern Recognition, 2019.

[4] Li D, Zhang Z, Chen X, et al. A richly annotated pedestrian dataset for person retrieval in real surveillance scenarios. IEEE Transactions on Image Process, 2018, 28(4):1575-1590.

[5] Huang K, Ren W, Tan T. A Review on image object classification and detection. Chinese Journal of Computers 2014, 37(6):1225-1240. (黄凯奇, 任伟强, 谭铁牛, 等. 图像物体分类与检测算法综述. 计 算机学报, 2014, 37(6):1225-1240.)

[6] Chen J, Chen Y, Li W. Application and Prospect of Deep Learning in Video Object Segmentation. Chinese Journal of Computers 2021, 44(3):609-631. (陈加, 陈亚松, 李伟浩, 等. 深度学习在视频对象分割中的应用与展 望. 计算机学报, 2021, 44(3):609-631.)

[7] Huang L, Zhao X, Huang K. Got-10k: A large high-diversity benchmark for generic object tracking in the wild. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019.

[8] Shan Y, Zhang Z, Huang K. Visual Human Action Recognition: History, Status and Prospects. Journal of Computer Research and Development. 2016, 53(1):93-112. (单言虎, 张彰, 黄凯奇. 人的视觉行为识别研究回顾, 现状及展望. 计算机研究与发展, 2016, 53(1):93-112.)

[9] Lecun Y, Bengio Y, Hinton G. Deep learning. Nature, 2015, 521(7553):436-444.

[10] Zhu J, Liao S, Lei Z, et al. Pedestrian attribute classification in surveillance: Database and evaluation//Proceedings of the IEEE International Conference on Computer Vision Workshops. Sydney, Australia, 2013: 331-338.

[11] Li D, Chen X, Huang K. Multi-attribute learning for pedestrian attribute recognition in surveillance scenarios//Proceedings of the IEEE Asia Conference on Pattern Recognition. Kuala Lumpur, Malaysia, 2015: 111-115.

[12] Li D, Zhang Z, Chen X, et al. A richly annotated dataset for pedestrian attribute recognition. arXiv preprint arXiv:1603.07054, 2016.

[13] Deng Y, Luo P, Loy C C, et al. Pedestrian attribute recognition at far distance//Proceedings of the ACM International Conference on Multimedia. 2014: 789-792.

[14] Liu X, Zhao H, Tian M, et al. Hydraplus-net: Attentive deep features for pedestrian analysis//Proceedings of the IEEE International Conference on Computer Vision. Orlando, USA. 2017: 350-359.

[15] Jia J, Huang H, Chen X, et al. Rethinking of pedestrian attribute recognition: A reliable evaluation under zero-shot pedestrian identity setting. arXiv preprint arXiv:2107.03576 2020.

[16] Li Q, Zhao X, He R, et al. Pedestrian attribute recognition by joint visual-semantic reasoning and knowledge distillation//Proceedings of the International Joint Conference on Artificial Intelligence. Stockholm, Sweden, 2018: 3177-3183.

[17] Wang J, Zhu X, Gong S, et al. Attribute recognition by joint recurrent learning of context and correlation//Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy, 2017: 531-540.

[18] Li Q, Zhao X, He R, et al. Recurrent prediction with spatio-temporal attention for crowd attribute recognition. IEEE Transactions on Circuits and Systems for Video Technology, 2019, 30(7):2167-2177.

[19] Tang C, Sheng L, Zhang Z, et al. Improving pedestrian attribute recognition with weakly-supervised multi-scale attribute-specific localization//Proceedings of the IEEE International Conference on Computer Vision. Seoul, Korea, 2019. 4997-5006

[20] Tan Z, Yang Y, Wan J, et al. Relation-aware pedestrian attribute recognition with graph convolutional networks//Proceedings of the

AAAI Conference on Artificial Intelligence, New York, USA, 2020,34: 12055-12062.

[21] Jia J, Chen X, Huang K. Spatial and semantic consistency regularizations for pedestrian attribute recognition// Proceedings of the IEEE/CVF International Conference on Computer Vision , Montreal, Canada, 2021:962-971

[22] Li Y, Huang C, Loy C C, et al. Human attribute recognition by deep hierarchical contexts//Proceedings of the European Conference on Computer Vision. Amsterdam, The Netherlands : Springer, 2016: 684-700.

[23] Wu M, Huang D, Guo Y, et al. Distraction-Aware Feature Learning for Human Attribute Recognition via Coarse-to-Fine Attention Mechanism//Proceedings of the AAAI Conference on Artificial Intelligence. New York, USA, 2020, 34(07): 12394-12401.

[24] Yaghoubi E, Khezeli F, Borza D, et al. Human attribute recognition—a comprehensive survey. Applied Sciences, 2020, 10(16): 5608.

[25] Wang X, Zheng S, Yang R, et al. Pedestrian attribute recognition: A survey. arXiv preprint arXiv:1901.07474, 2019.

[26] Kang Y, Xie Y, Zhang S, et al. Research Progress and Application Exploration of Key Technologies for Human Image Attribute Recognition. Police Technology, 2018(2):12-16 (康运锋, 谢元涛, 张世渝. 人像属性识别关键技术研究进展及应用探索. 警察技术, 2018(2):12-16.)

[27] Guo H, Zheng K, Fan X, et al. Visual attention consistency under image transforms for multi-label image classification//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 729-739.

[28] Lin T Y, Maire M, Belongie S, et al. Microsoft coco: Common objects in context//Proceedings of the European Conference on Computer Vision. Zurich, Switzerland: Springer, 2014: 740-755.

[29] Everingham M, Van Gool L, Williams C K, et al. The pascal visual object classes (voc) challenge. International Journal of Computer Vision, 2010, 88(2):303-338.

[30] Kuznetsova A, Rom H, Alldrin N, et al. The open images dataset v4. International Journal of Computer Vision, 2020, 128(7): 1956-1981..

[31] Deng J, Dong W, Socher R, et al. Imagenet: A large-scale hierarchical image database//2009 IEEE conference on computer vision and pattern recognition. Miami, USA. 2009: 248-255.

[32] Krizhevsky A, Hinton G, et al. Learning multiple layers of features from tiny images. 2009.

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

[34] Tan Z, Yang Y, Wan J, et al. Attention-based pedestrian attribute analysis. IEEE Transactions on Image Process, 2019, 28(12):61266140.

[35] Wu J, Liu H, Jiang J, et al. Person attribute recognition by sequence contextual relation learning. IEEE Transactions on Circuits and Systems for Video Technology, 2020, 30(10):3398-3412.

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

[37] Bourdev L, Maji S, Malik J. Describing people: A poselet-based approach to attribute classification//Proceedings of the IEEE International Conference on Computer Vision. Barcelona, Spain. 2011: 1543-1550.

[38] Sharma G, Jurie F. Learning discriminative spatial representation for image classification//Proceedings of the British Machine Vision Conference. Dundee, UK , 2011: 1-11.

[39] Zheng L, Shen L, Tian L, et al. Scalable person re-identification:A benchmark//Proceedings of the IEEE International Conference on Computer Vision. Santiago, Chile , 2015: 1116-1124.

[40] Ristani E, Solera F, Zou R, et al. Performance measures and a data set for multi-target, multi-camera tracking//Proceedings of the European Conference on Computer Vision. Amsterdam, Netherlands: Springer, 2016: 17-35.

[41] Geiger A, Lenz P, Urtasun R. Are we ready for autonomous driving? the kitti vision benchmark suite//2012 IEEE Conference on Computer Vision and Pattern Recognition. Rhode Island,USA, 2012: 33543361.

[42] Bileschi S M. Streetscenes: Towards scene understanding in still images. Massachusetts Inst of Tech Cambridge,USA 2006.

[43] Dalal N, Triggs B. Histograms of oriented gradients for human detection//2005 IEEE computer society conference on computer vision and pattern recognition (CVPR2005): Boston, USA, 2005,1: 886-893.

[44] Bourdev L, Malik J. Poselets: Body part detectors trained using 3d human pose annotations//Proceedings of the IEEE International Conference on Computer Vision. Kyoto, Japan, 2009: 1365-1372.

[45] Maji S. Large scale image annotations on amazon mechanical turk:UCB/EECS-2011-79. EECS Department, University of Califor-nia,Berkeley,2011.http://www.eecs.berkeley.edu/Pubs/TechRpts/2011/EECS-2011-79.html.

[46] Felzenszwalb P F, Girshick R B, Mcallester D, et al. Object detection with discriminatively trained part-based models. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2009, 32(9):1627-1645.

[47] Xiong Y, Zhu K, Lin D, et al. Recognize complex events from static images by fusing deep channels//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA, 2015: 1600-1609.

[48] Yang Z, Kai Y, Biao L, et al. Weakly-supervised learning of midlevel features for pedestrian attribute recognition and localization// Proceedings of the British Machine Vision Conference. London, UK, 2017: 69.1-69.12.

[49] Sarfraz M S, Schumann A, Wang Y, et al. Deep view-sensitive pedestrian attribute inference in an end-to-end model//Proceedings of the British Machine Vision Conference. London, UK, 2017. 70.1-70.12.

[50] Liu P, Liu X, Yan J, et al. Localization guided learning for pedestrian

attribute recognition//Proceedings of the British Machine Vision Conference. Newcastle, UK , 2018.

[51] Li D, Chen X, Zhang Z, et al. Pose guided deep model for pedestrian attribute recognition in surveillance scenarios//Proceedings of the IEEE International Conference on Multimedia & Expo. San Diego, USA, 2018: 1-6.

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

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

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

[55] Li Q, Zhao X, He R, et al. Pedestrian attribute recognition by joint visual-semantic reasoning and knowledge distillation//Proceedings of the International Joint Conference on Artificial Intelligence. Macao, China, 2019: 833-839.

[56] Han K, Wang Y, Shu H, et al. Attribute aware pooling for pedestrian attribute recognition//Proceedings of the International Joint Conference on Artificial Intelligence. Macao, China , 2019.

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

[58] Zeng H, Ai H, Zhuang Z, et al. Multi-task learning via co-attentive sharing for pedestrian attribute recognition//2020 IEEE International Conference on Multimedia and Expo (ICME). London, UK, 2020: 1-6.

[59] Zhang J, Ren P, Li J. Deep template matching for pedestrian attribute recognition with the auxiliary supervision of attribute-wise keypoints. arXiv preprint arXiv:2011.06798, 2020.

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

[61] He K, Zhang X, Ren S, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(9):1904-1916.

[62] Pascanu R, Mikolov T, Bengio Y. On the difficulty of training recurrent neural networks//International conference on machine learning. Atlanta, USA: PMLR, 2013: 1310-1318.

[63] Hochreiter S, Schmidhuber J. Long short-term memory. Neural computation, 1997, 9(8):1735-1780.

[64] Zitnick C L, Dollár P. Edge boxes: Locating object proposals from edges//Proceedings of the European Conference on Computer Vision. Zurich, Switzerland : Springer, 2014: 391-405.

[65] Wei S E, Ramakrishna V, Kanade T, et al. Convolutional pose machines//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA , 2016: 4724-4732.

[66] Jaderberg M, Simonyan K, Zisserman A, et al. Spatial transformer networks//Advances in Neural Information Processing Systems. Montreal, Canada, 2015: 2017-2025.

[67] Zhao H, Tian M, Sun S, et al. Spindle net: Person re-identification with human body region guided feature decomposition and fusion// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2017: 1077-1085.

[68] Lin T Y, Goyal P, Girshick R, et al. Focal loss for dense object detection//Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy, 2017: 2980-2988.

[69] Xingjian S, Chen Z, Wang H, et al. Convolutional lstm network: A machine learning approach for precipitation nowcasting//Advances in neural information processing systems. Montreal, Canada, 2015: 802-810.

[70] Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531, 2015.

[71] Kalayeh M M, Basaran E, Gökmen M, et al. Human semantic parsing for person re-identification//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 1062-1071.

[72] Lin T Y, Dollár P, Girshick R, et al. Feature pyramid networks for object detection//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2017: 2117-2125.

[73] Arandjelovic R, Gronat P, Torii A, et al. Netvlad: Cnn architecture for weakly supervised place recognition//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 5297-5307.

[74] Zhou X, Wang D, Krähenbühl P. Objects as points. arXiv preprint arXiv:1904.07850, 2019.

[75] Gray D, Brennan S, Tao H. Evaluating appearance models for recognition, reacquisition, and tracking//Proceedings of the IEEE international workshop on performance evaluation for tracking and surveillance: Beijing, China , 2007,3: 1-7.

[76] Zhou B, Khosla A, Lapedriza A, et al. Learning deep features for discriminative localization//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 29212929.

[77] Girshick R. Fast r-cnn//Proceedings of the IEEE International Conference on Computer Vision. Santiago, Chile , 2015: 1440-1448.

[78] Kipf T N, Welling M. Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907, 2016.

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

[80] Graves A, Fernández S, Gomez F, et al. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks//Proceedings of the 23rd international conference on Machine learning. Pittsburgh, USA , 2006: 369-376.

[81] Ioffe S, Szegedy C. Batch normalization: Accelerating deep network

training by reducing internal covariate shift//International conference on machine learning. Lille, France: PMLR, 2015: 448-456.

[82] Szegedy C, Vanhoucke V, Ioffe S, et al. Rethinking the inception architecture for computer vision//Proceedings of the IEEE conference on computer vision and pattern recognition. Las Vegas, USA, 2016: 2818-2826.

[83] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA , 2016: 770-778.

[84] Gkioxari G, Girshick R, Malik J. Contextual action recognition with r* cnn//Proceedings of the IEEE International Conference on Computer Vision. Santiago, Chile , 2015: 1080-1088.

[85] Guo H, Fan X, Wang S. Human attribute recognition by refining attention heat map. Pattern Recognition Letters, 2017, 94:38-45.

[86] Zhu F, Li H, Ouyang W, et al. Learning spatial regularization with image-level supervisions for multi-label image classification// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2017: 5513-5522.

[87] Yang J, Fan J, Wang Y, et al. Hierarchical feature embedding for attribute recognition//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 13055-13064.

Jia Jian, Ph.D. candidate. His research in computer vision, pattern recognition, pedestrian attribute recognition,

Chen Xiaotang, Ph.D., associate professors. Her research in computer vision, pattern recognition.

Huang Kaiqi, Ph.D., professors. His research interests include computer vision, pattern recognition, visual surveillance，and cognitive decision.

Background

Pedestrian attribute recognition has received a lot of attention due to its wide range of applications. As a critical component, it can be widely used in person re-identification, person search, and person retrieval. Over the past few years, thanks to the boom in deep learning, a variety of methods and dataset have been proposed to facilitate the development of pedestrian attribute recognition. Therefore, from the perspective of task definition and conceptual scope, this paper overviews the recent works and progress in pedestrian attribute recognition. In addition, in terms of application scenario and method category, we compare pedestrian attribute recognition with the two related tasks, human attribute recognition and pedestrian attribute recognition oriented to pedestrian re-identification, to present the essential characteristic of pedestrian attribute recognition directly.

Pedestrian attribute recognition is the main research topic in the Center for Research on Intelligent System and Engineering, Institute of Automation, Chinese Academic of Science. We researched the robustness and consistency regularizations of pedestrian attribute recognition in the last three years and published three related papers.

This survey is supported in part by the National Natural Science Foundation of China (Grant No. 61721004), the Projects of Chinese Academy of Science (Grant No. QYZDB-SSW-JSC006), and the Strategic Priority Research Program of Chinese Academy of Sciences (Grant No. XDA27000000).