分类号 _____ 密级 _____

UDC[注 1] _____

南京理工大学

NANJING UNIVERSITY OF SCIENCE & TECHNOLOGY

博 士 学 位 论 文

面向不同监督条件下的行人再识别关键问题研究

（题名和副题名）

丁国栋

（作者姓名）

指导教师姓名 白光一 教授

唐振民 教授

学 位 类 别 工学博士

学 科 名 称 计算机科学与技术

研 究 方 向 图像分析与检索

论文提交日期 **2019.12**

注 1: 注明《国际十进分类法 UDC》的类号

面向不同监督条件下的行人再识别关键问题研究

南京理工大学

博 士 学 位 论 文

面向不同监督条件下的行人再识别关键问题研究

作　　者：丁国栋

指导教师：白光一 教授

唐振民 教授

南京理工大学

2019 年 12 月

Ph.D. Dissertation

Research on Person Re-identification Methods under Different Supervisions

By

Guodong Ding

Supervised by Prof. ***Guangyi Bai***

Zhenmin Tang

Nanjing University of Science & Technology

December, 2019

声　明

本学位论文是我在导师的指导下取得的研究成果, 尽我所知, 在本学位论文中, 除了加以标注和致谢的部分外, 不包含其他人已经发表或公布过的研究成果, 也不包含我为获得任何教育机构的学位或学历而使用过的材料。与我一同工作的同事对本学位论文做出的贡献均已在论文中作了明确的说明。

研究生签名: _____　　年　　月　　日

学位论文使用授权说明

南京理工大学有权保存本学位论文的电子和纸质文档, 可以借阅或上网公布本学位论文的部分或全部内容, 可以向有关部门或机构送交并授权其保存、借阅或上网公布本学位论文的部分或全部内容。对于保密论文, 按保密的有关规定和程序处理。

研究生签名: _____　　年　　月　　日

摘 要

深度学习方法在自然图像、自然语言等众多机器学习领域取得了较好的性能，从而引起了学术界和工业界的广泛关注。深度学习研究的是一大类拥有多层非线性变换的模式识别系统，这些系统将数据从低往高逐层抽象，相比于浅层方法更适合于对真实世界中的高度非线性问题进行建模。

随着人们对社会公共安全的日益关注以及视频采集技术和大规模数据存储技术的发展，我们对于大规模监控系统下的视频内容进行自动化和智能化分析的需求也越来越多。而行人再识别则属于智能监控领域的一个重要研究方向，其主要研究在监控视频中识别出某个特定的已经在监控网络中出现过的行人。该问题的研究能够为智能监控领域的诸多应用提供基础技术支撑和可行性保障，例如，辅助行人跟踪。

本文围绕行人再识别问题，通过利用深度神经网络模型，从不同的数据约束条件下出发，即：从全监督到半监督，再到无监督的情况下，分别进行了不同层面的研究探索，分别采用卷积神经网络、模型正则化以及聚类等机器学习技术，开展了若干关键技术的研究工作。本文具体的研究内容以及创新之处主要包含以下几个方面：

(1) 在全监督设定下，针对单支网络行人特征提取的欠完备问题，本文提出了一种互补鉴别特征提取算法。通过设计一个类孪生特征提取网络架构，利用特征掩码模块连接主网络以及分支网络，以主网络的特征作为分支网络输入的一部分，在其基础上进行更多的互补特征的挖掘。随后，在此基础上加入了跨支成对排序损失函数进行网络训练，使得生成的分支网络特征更加具有互补性。实验数值以及可视化表示的结果均表明，所提出的方法显著提升了行人再识别的准确率。

(2) 在半监督设定下，针对行人再识别数据库中有标注行人数量少，深度学习网络参数在训练中会出现的“过拟合”的问题，本文提出了利用生成对抗网络生成伪数据并为伪数据生成伪标签的方法来进行模型的正则化。通过在行人特征表示空间中，利用无标签样本与有标签的行人样本特征表示之间的相似性为其分配伪标签来辅助训练。基于特征表示的伪标签生成准则能够在统一的框架下生成两种标签编码，分别是独热码和分布码。在公开的数据集上进行的实验取得了较高的行人再识别性能并证明了该方法的有效性。

(3) 在无监督设定下，针对现有的行人再识别聚类方法中的簇候选和合并准则无法达到较好的聚类效果的问题，本文提出了使用分散度作为簇候选和合并的标准。同时衡量和考虑簇间分散度和簇内分散度能够优化聚类的结果，从而最终取得更好的行人再识别结果。基于分散度的准则不仅能够自动为孤立的数据点提升聚类优先级，还能够阻止坏聚类的形成和蔓延。除此之外，该分散度的合并准则与卷积神经网络之间能够起到相互促进的作用，拥有更快的收敛速度和更好的稳定性。该

方法在基于图像和视频的公开数据库上都取得了领先的无监督行人再识别性能。

关键词： 行人再识别，卷积神经网络，互补特征，伪标签，聚类

Abstract

Deep learning approaches have demonstrated their capability in resolving multiple machine learning tasks, such as image processing and natural language processing, and achieved state-of-the-art performances, which draws intense attention from both academia and industry. Deep learning studies a large range of pattern recognition systems with multi-layer nonlinear transformations. These systems abstract data from low to high levels, and are more suitable for highly nonlinear problems in the real world than the shallow approaches with less layers.

With the increasing attention to social public safety and the development of video capture technology and large-scale data storage technology, there is an increasing demand for automated and intelligent analysis of video content under large-scale surveillance systems. Person re-identification is an important research topic in the field of intelligent monitoring. It mainly studies the identification of a specific person in the surveillance video who has already appeared in the monitoring network. The research of this problem can provide basic technical support and feasibility guarantee for many applications in the field of intelligent monitoring, for example, pedestrian tracking.

This dissertation focuses on the issue of person re-identification, works under different supervision constraints, from full supervision to semi-supervised and then unsupervised, and carries out different levels of research and exploration. Multiple machine learning techniques such as convolutional neural networks (CNN), model regularization technique and clustering have been studied and researched in this work. The specific research content and innovations of this dissertation mainly include the following aspects:

(1) Under the setting of full supervision, this dissertation proposes a complementary discriminative feature extraction algorithm to deal with the shortcoming of losing local features by the single-branch network. By designing a Siamese-like feature extraction network architecture, the feature mask module is used to connect the main branch and the other branch, thus features from the main branch can be used as a part of input to the other branch's for complementary feature mining. On the basis of this, an inter-branch pairwise ranking loss function is added on top of the network for training, which promotes the diversity and complementary characteristics between two branches. Experimental numerical results and visual representations show that the proposed method significantly improves the accuracy of person re-identification.

(2) Under the setting of semi-supervision, in order to combat the problem of "over-fitting" in the person re-identification task whose datasets only contain limited numbers of annotations while deep learning network are data-hungry. This dissertation proposes a method for generating pseudo-labels for pseudo-data to perform model regularization. A pseudo-label is assigned to an unlabeled sample according to its similarity to the labeled person representations in the feature space. The pseudo-label generation criterion based on the feature representation can provide two schemes of label encoding under the unified framework, i.e., the one-hot scheme and the distributed scheme. Experiments were conducted on public datasets to demonstrate its competitive performance to its counterparts.

(3) Under the setting of no supervision, the clustering and merging criterion in the existing person re-identification clustering methods can not achieve satisfactory clustering results. This dissertation proposes to use dispersion as a cluster selection and merging criterion. By simultaneously measuring the inter-cluster dispersion and intra-cluster dispersion, the results of clustering can be optimized to boost person re-identification performances. The criterion based on dispersion not only can automatically raise the clustering priority for isolated points, but also prevents the formation of bad clusters. In addition, the method can play a mutually reciprocal promotion role with the convolutional neural network, leading to faster convergence and better stability. State-of-the-art person re-identification performance has been achieved on both images and video based datasets.

Keywords: Person Re-Identification, Convolutional Neural Network, Complementary Representation, Pseudo-Labeling, Clustering

目录

图表目录

1 绪论

1.1 课题的研究背景及意义

随着社会和经济的飞速发展，人们对于交通、安检、银行、军事等不同领域的安全防范要求不断提高。得益于网络信息技术和智能设备制造业的快速发展，大规模的摄像头布置现在在大城市中是随处可见的，这些摄像头每天都会产生海量的视频监控数据。然而拥有大量数据，没有机器的自主学习分析和处理归纳能力，并不意味着能够获得海量信息。同样，拥有海量信息，没有机器的自主学习分析和处理归纳能力，也不代表能够掌握丰富知识。因此，提升机器对监控视频数据的分析、处理、学习和归纳能力是极其重要的。而目前的视频监控技术主要以“人工分析”为主，结合简单的智能化方法来处理分析视频数据，需要耗费大量的人力成本，同时也带来了诸如：“视频在、找不到”，“找得到、找太久”，“有服务、不可靠”等视频监控技术应用瓶颈。因此，智能视频技术的发展程度也决定了大规模监控系统下，对视频内容进行自动化、智能化理解分析水平的高低，为信息获取的有效性以及后续管理者决断的及时性提供了技术保障和支撑。由此可见，智能视频技术的发展已经变得尤为重要。智能视频技术的发展往往依赖于计算机视觉处理技术的发展状况。此项技术对视频图像信号进行自动化处理、关键信息获取以及语义内容的理解，挖掘视频中的异常情况，为工作人员及时地提供有效信息。目前比较典型的应用包括了“运动目标检测、目标跟踪、目标识别”等实际应用。

行人目标是监控视频里常见的和比较受关注的目标，能够实现对视频监控数据中出现的行人进行识别和追踪，将能为公安部门的案件侦破提供极大的便利。而如何识别出某个特定的已经在监控网络中出现过的行人成为了智能监控领域的一个重要需求，该研究方向被定义为行人再识别，行人再识别的研究能够为智能视频监控的其他诸多应用的解决方案提供基础的技术支撑和可行性保障。

行人再识别，是利用计算机视觉技术判断不同视角下无交叉的图像或者视频序列中是否存在特定行人的技术。行人再识别来源于行人跟踪问题[1,2]，其研究背景主要是基于大规模无交叉的摄像机监控系统。在现实生活中的大多数应用场景中，由于行人在运动过程中，一段时间内会从一个视角的摄像机中消失，并出现在另一个视角摄像机的监控下，或者经过一段时间后，再次出现在同一个视角中，因此，研究如何针对大规模无交叉视角的监控系统，有效地解决目标的自动识别与定位，促进目标检测与跟踪的准确性成为计算机视觉领域的主要研究问题之一。人脸对于行人身份的识别是有效的，然而在开放式的监控场景下，精细的人脸信息的获取并不能得到保证，其主要原因在于以下两个方面：一方面，一般的监控摄像头安装位置距离受采集的行人较远，所获取的人脸部分分辨率比较低，从而很难从其中提取出具有鉴别信息的人脸特征；另一方面，监控

摄像头的安装位置是固定的，而行人的行径方向却无法预估，从而导致采集到的行人图像有很大概率是侧面或者背面，而这些图片本身就不包含有效的人脸信息。考虑到在短时间内，行人在经过不同摄像头下，其外观（服装、背包、携带物品等）并不会发生剧烈的变化，在这样的情况下，通过行人的外观来完成行人再识别的任务则更加可行。

行人再识别问题的一般研究流程可以包括行人目标确定、行人目标特征表示（人体模型、特征描述符等）以及行人特征的相似性度量等几个核心部分，具体如图1.1所示。在实际应用中，当从监控摄像头获取得到的监控视频流输入的时候，首先第一步需要执行的是行人检测的步骤，通过使用一些人体模型或者深度学习的行人检测算法将行人框出来，在得到行人的标注框之后的步骤即为行人的特征提取，并最终使用该行人的特征表示在待查询集中进行相似度计算，从而找到最为匹配的行人作为最终的识别结果。

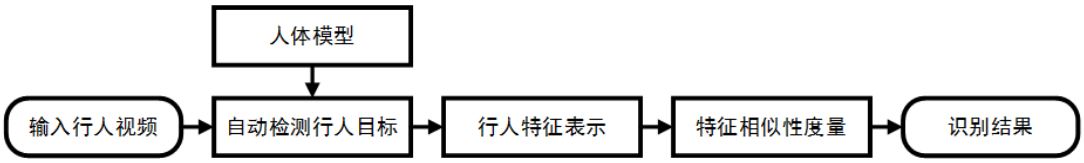


图 1.1 行人再识别整体框架图

如图1.1所示，行人再识别主要关注的点是行人目标检测之后的流程，所以其技术难点也就在行人特征表示和特征相似性度量这两大块。然而解决这一课题也同样面临着诸多挑战。这些挑战可以归纳为两种：第一个挑战是行人视觉表观差异性很大；第二个挑战是模型的训练对数据量有很大的需求。本文主要的研究内容就在于如何在有限的数据约束下进行良好的行人特征表示的学习，下面本章将分别具体介绍一下这两种挑战出现的原因以及他们的表现形式。

首先，行人视觉表观差异性大主要表现在以下几个方面：

- 行人本身差异性大，首先行人人种、身形、体态各不相同（有小孩、老人等）；行人衣服穿着千差万别（不同颜色、材质等）。
- 不同监控设备受到自身参数、性能、状态等多方面的限制，且在大规模视频监控系统中，由于成本等问题，摄像机的分辨率往往较低。这也直接导致了采集到的视频行人数据大多模糊不清，一些基于细节的算法，例如：人脸识别等方法，也难以应用其中。
- 监控摄像头安装位置不同。由于摄像头安装的位置固定，导致了其视角的固定，相邻两个摄像头的安装位置，角度都不尽相同，在拍摄参数不相同的情况下即使是同一个行人，采集到的图像也会有可能出现很大的差异。
- 行人行径路线不同。行人可能会从各个方向进入到监控视野中，从而不能保证行人被捕捉到时所处的状态，比如说，行人被拍到的是正面，背面还是侧面都是未知的。
- 行人在摄像头中的位置不固定，可能在中途的一些时间点存在被障碍物或其他行人遮挡的可能，而这些遮挡则是导致行人再识别性能较低一个比较重要的原因。

其次，利用深度神经网络方法来进行行人再识别需要大量带标注的训练数据。而训练数据的获取和标注也面临着很大的挑战，这些挑战主要表现在以下三个方面：

- 训练数据数量有限。从当前行人再识别训练数据的收集情况来看，收集到的数据相对于真实数据的时空分布是非常有限的、局部的。同时，与其他视觉任务相比，行人再识别的数据规模也是非常小的。比如以大规模图像识别数据集 ImageNet 来说，它的训练数据有 125 万张图片，在行人检测数据集[3] 上标注的行人框有 35 万个，COCO 的目标检测数据训练集是 12.3 万多张图片。而在行人再识别领域中，目前常用的数据集仅有 3 万多张行人图片。相比之下还是有比较大的差距。
- 数据采集困难。正如第一个挑战所述，行人再识别的图像中的行人视觉表现差异性很大，那么如何采集到这些包含多种多样的行人变化的数据本身也是一种挑战。从现实的角度来说，完备的数据采集是不可能实现的，因为采集的变量变化非常复杂，例如，相机的选择、安装位置的选择、拍摄时天气、光照等等都会起到不同的影响。除此之外，被拍摄行人的隐私问题也对数据采集和获取造成了阻碍。
- 数据标注比较困难。首先大量的数据标注本身就需要浩大的工作量，例如，大规模图像分类数据集 ImageNet，通过众包的形式前后有 4.8 万人花了近两年时间来标注。无论从时间还是金钱上来看，标注成本都是非常高的。其次，针对行人再识别的标注本身也是非常困难的，相比于分类的标注中简单地把狗和猫分开，在视频中把两个年龄、体貌相似，穿着类似衣服的不同行人分开是更加困难的。

以上的种种原因，造成了行人在监控系统中所采集到的图像数据较少、行人在外观上有较大差异，使得利用深度神经网络进行行人再识别任务面临着非常大的挑战。

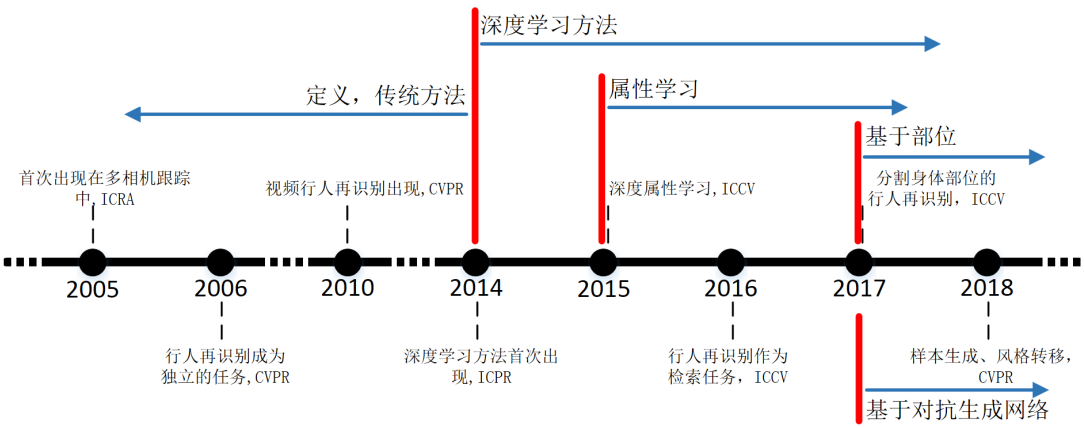


图 1.2 行人再识别领域发展历程里程碑

行人再识别领域发展历程以及里程碑式的工作，按照时间顺序在图1.2中作了展示[1]。在 2005 年，“行人再识别”的概念首次出现在了多摄像头跟踪的工作中[5]。2006 年，Gheissari 等人[6] 将“行人再识别”从“多摄像头跟踪”中分离出来，并将其作为一个独立的计算机视觉任务来进行研究。2010 年，基于视频的行人再识别问题在 CVPR 中

[1] 图片参考自文献 [4]

被提出，将从单幅图像的行人再识别推广到更多图像帧的情况，行人之间的相似性通过综合这一系列的图像帧集合之间的相似性来得到。2014 年，深度学习在图像分类任务上取得了巨大的成功，因此深度学习技术也被广泛地应用到了行人再识别领域，具体来讲，卷积神经网络被用到了行人的特征提取上来。在随后的几年内，越来越多的研究分支被广泛挖掘。2012 年，Liu 等人[7] 提出了深度属性学习，其主要思想是通过使用一些行人身上的特定的属性来辅助行人特征的学习，例如，长发、白色短袖，背包等等。这些属性作为一些额外的信息，可以优化行人再识别的特征提取以及提升特征的鲁棒性。2015 年，Zheng 等人[8] 又重新将行人再识别作为一个检索问题来进行研究，启发了一系列与检索相关的工作。2017 年，随着生成对抗网络 (Generative Adversarial Network, GAN)[9] 的出现，很多工作利用生成对抗网络来生成更多的训练样本来正则化训练模型。这样的工作在一定程度上减轻了模型训练对于数据量的需求，因为其可以被看做是一种数据增广的方法。与此同时，也有着大量工作通过利用更为细致的身体切割来学习不同部位的特征，从而得到更好的行人表示来进行行人再识别。这些工作的提出进一步地将行人再识别往更细粒度的领域推进。

行人再识别领域所受到的关注度还可以从计算机视觉领域顶级会议的稿件接收数量中窥探一二。图1.3统计了近 6 年来在计算机视觉领域三大顶级会议（CVPR，ICCV 和 ECCV）上出现行人再识别相关的接收论文数量，可见其迅猛增长之趋势。

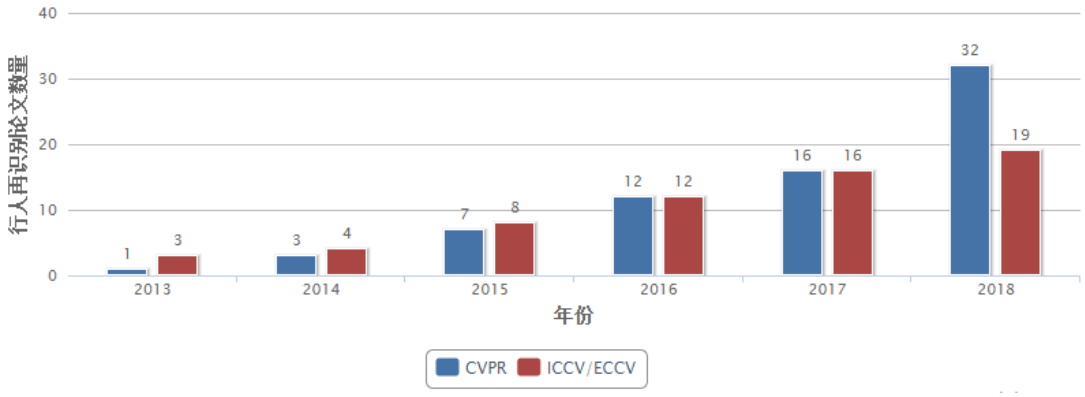


图 1.3 近六年计算机视觉顶级会议中行人再识别相关论文的数量

行人再识别作为计算机视觉领域自成一派的研究方向，与其他同等受到广泛关注的图像分类和图像检索有着密不可分又独树一帜的关联。根据训练集和测试集之间的关系，行人再识别可以看成一个处于图像分类[10] 和目标检索[11] 之间的任务（表1.1）。对于图像分类任务而言，每个类别都有训练图像，并且测试图像属于这些预定义的类，如表1.1中“可见”所示。对于目标检索任务而言，通常没有训练数据，因为查询样本本身是预先不可知的，同样地、检索目标图库可能包含各种类型的对象。因此，训练类别是“不提供的”，测试类别（查询）同样也是“不可见”的。与图像分类相比，行人再识别与之相似之处在于训练类别可见，即包含了不同身份的行人图像，相同身份的图像属于同一类。行人再识别同样也类似于目标检索，都需要对一个检索图像进行检索，并

且测试身份是不可见的，它们与训练集中的身份也无任何重叠，除了训练和测试图像都是行人罢了。所以，与图像分类任务相比，行人再识别的测试类别是未知的；与图像检索任务相比，行人再识别的测试类别也是未知的。但正因为这样，行人再识别任务可以通过综合考虑和利用分类和检索方面的技术来促进解决方案的提出。一方面，可见训练类别的提供使得模型可以在特征空间中学到良好的距离度量[12] 或特征表示[13,14]。另一方面，当涉及到检索时，有效的索引结构[15] 和哈希技术[16] 可以有益于大型图库中的行人再识别任务的快速进行。

表 1.1 行人再识别任务与图像分类和图像检索的对比

任务	训练类别	测试类别	目标
图像分类	提供	可见	鉴别性学习
图像检索	未提供	不可见	高效
行人再识别	提供	不可见	鉴别学习 + 高效

行人再识别最终任务在于在候选集（Gallery）中找出与查询集（Probe）中身份相匹配的行人。目前的测试过程可分为两种模式：第一种模式叫做“验证”(verification)，即：对于一个待查询行人样本，分别与行人数据库中的每一张组成对 (pair) 并对其是否属于同一身份进行判别，其本质是个“二分类”的问题。这种模式的缺点在于当数据集规模较大时，相应所组成的样本对会成倍增长，会导致这样的验证方法效率的降低；此外，这种验证模式也忽略了不同行人之间的差异性而仅仅关注输入是否具有相同的行人身份。第二种模式叫做“检索”（retrieval），其是借鉴于图像检索，通过在训练集上得出的能够提取身份鉴别信息的模型，在测试集上进行特征提取，并基于这些提取出来的特征进行距离排序，最终利用得到排序列表来评估其性能。由于后者具有更强的可操作性以及对大数据集更好的泛化能力，在业界被广泛采用。本文中的行人再识别测试过程也采用第二种（检索）模式。

1.2 行人再识别研究现状

针对行人再识别问题面临的多种挑战，计算机视觉领域专家、学者在过去几年时间内做了大量的研究工作，提出了丰富的模型和方法。国外的研究机构主要有英国的伦敦玛丽女王大学（QMUL）、澳大利亚的悉尼大学（Sydney University, SYU）、悉尼科技大学（UTS）、新加坡的南洋理工大学（National Nanyang University, NTU）、美国的东北大学 (Northeastern University)、德克萨斯大学（University of Texas）等。其中，英国伦敦玛丽女王大学的 Shaogang Gong 教授是最早一批研究行人再识别问题的，他们的团队在经典方法以及深度学习方法都有着很多典型的工作[17–21]，对行人再识别领域的发展产生了意义深远的影响。悉尼科技大学的 Yi Yang 同样也是产出了很多高质量的工作[22–25]，性能和创新性都在业界是比较领先的。美国德克萨斯大学的 Qi Tian 教

授同样也提出了很多有着领先性能的行人再识别模型[26–29]。而国内的研究机构主要有清华大学、香港中文大学、中山大学、北京大学、浙江大学、厦门大学、大连理工大学等。其中清华大学毕业生 Liang Zheng 在行人再识别领域作出了很多重要工作，包括提出了多个行人再识别数据库[8,14]。相似地，香港中文大学也提出了一系列的行人再识别数据集[13]。中山大学的 Weishi Zheng 教授与 Shaogang Gong 教授有着密切合作，同样也一直在行人再识别领域做着持续的贡献[30]。

总的来看，对于大规模视频监控网络下的行人再识别问题的研究方法主要可以概括为两类，分别是：基于特征表示的行人再识别方法和基于度量学习的行人再识别方法，下面将分别介绍这两方面工作的研究进展。

1.2.1 基于特征表示的行人再识别方法

基于特征表示的行人再识别方法的主要目标就是优化行人特征提取器，使得其提取出来的行人特征能够更加具有鲁棒性。该类方法根据是否使用深度卷积神经网络学习技术分为经典方法和深度方法，其具体的一些介绍如下：

经典方法 在传统的行人特征表示提取中，最常被使用到的特征往往是一些比较低层次的特征，比如说颜色特征，相对来说，纹理特征则使用的较少。Gheissari 等人[6] 在 2006 年提出了一种空间-时间 (spatial-temporal) 的分割方法来检测稳定的前景区域。对于一个局部区域，计算其 HS 直方图和边缘直方图。后者捕捉到的是起主导作用的局部边界方向和边缘两侧的 RGB 比率。Gray 和 Tao[31] 在亮度通道上采用了 8 个颜色通道滤镜 (RGB，HS 和 YCbCr) 和 21 个纹理滤镜，并将行人划分为水平条纹。2010 年，在文献[32]中，Farenzena 等人提出将行人作为关心目标的前景从背景中分离出来，并为身体的每个部位计算对称轴。基于身体的拓扑结构信息，计算加权颜色直方图（Weighted color Histogram，WH），最大稳定颜色区域 (maximally stable color regions, MSCR) 和循环高结构化块（recurrent high-structured patches, RHSP）。其中，加权颜色直方图根据像素与对称轴距离的远近为其分配不同的权重，距离越近，权重越大，相反，距离越远，权重越小；接着还为每个身体部分构建颜色直方图。最大稳定颜色区域检测稳定的颜色区域并提取颜色，区域和质心等特征。循环高结构化补丁是一种捕获循环纹理块中所包含的纹理特征。后来许多的工作，例如文献[33]，文献[34]和文献[35]都采用的是与文献[31]相同的一组特征。类似地，Mignon 等人[36] 通过从 RGB，YUV 和 HSV 通道构建特征向量，并在水平条纹中构建 LBP 纹理直方图。

与上述早期的工作相比，近年来手工特征的选择基本保持不变[12,37–40]。在 Zhao 等人的一系列工作[38,41,42] 中，首先通过对图像进行步长为 5，长宽为 10 的致密图像块采样，接着从中提取出 32 维的 LAB 颜色直方图和 128 维的 SIFT 描述符作为特征表示；文献[43]中也同样采用了这样的图像块特征，但引入了邻接约束搜索用于在图库图像中具有相似高度的水平条纹中搜索查找与查询图像块相似的最佳匹配块。2013 年，Das 等人[44] 在文献[45]中提出的轮廓的基础上分别提取头部，躯干和腿部的 HSV 直方图。Li 等人[39] 也同样地从图像块中提取局部颜色描述符，但是使用分层高斯化[46] 聚合它们

来捕获空间信息，后续的文献[47]也采用相似的方式。Pedagadi 等人[48] 则首先从 HSV 和 YUV 空间提取颜色直方图和矩然后再使用 PCA 进行降维。Liu 等人[49] 对每个图像块分别进行了 HSV 直方图，梯度直方图和 LBP 直方图的提取。为了提高 RGB 值对光度方差的鲁棒性，Yang 等人[50] 在 2014 年引入了基于显著颜色名称的颜色描述符(SCNCD)，用于全局行人颜色描述；同时还分析了背景和不同颜色空间的对其的影响。在文献[12]中，Liao 等人提出局部最大出现（LOMO）描述符，其包括颜色和 SILTP 直方图。具体方法是对同一水平条纹中的块区域进行最大池化操作，构建三层金字塔模型并且在最后进行对数转换。LOMO 后来被文献[19]和文献[51]所使用，Chen 等人[40] 也使用了一组类似的特征。2015 年，在文献[8]中，Zheng 等人提出为每个局部图像块提取出 11 维的颜色名称描述符[52]，并通过词袋模型（BoW）模型将它们聚合成一个全局向量。2016 年，在文献[53]中，Matsukawa 等人提出了层次高斯特征来描述颜色和纹理线索，其通过多个高斯分布对每个区域进行建模。每个分布代表了区域内的图像块。

除了直接使用一些比较低层次的颜色和纹理特征之外，还有一些工作使用了基于属性的特征，可以将其视为中级表示。与低级描述符相比较，属性对图像的变换来说具有更加稳健的特性。在文献[54]中，Layne 等人在 VIPeR 数据集上标注了 15 个二进制属性分别表示一些不同的服装和生物识别属性。低级的颜色和纹理特征则用于训练属性分类器。在进行属性加权之后，结果向量被放到 SDALF[32] 框架中与其他视觉特征进行融合从而作为最终的表示。Liu 等人[7] 使用标注过的属性来改进 Latent Dirichlet Allocation（LDA）模型，以滤除掉嘈杂的 LDA 信号。Liu 等人[18] 则提出以给予一些共同行人属性来进行无监督的行人原型挖掘，并根据原型自适应地确定不同查询人的特征权重。最近的一些工作借助于其他数据进行属性学习。2015 年，在文献[55]中，Su 等人将同一个人但不同摄像机的二进制语义属性嵌入到连续的低秩属性空间中，使得属性向量对于匹配更具辨别力。Shi 等人[56] 建议可以从现有的时尚摄影数据集中学习许多属性，包括颜色，纹理和类别标签。这些属性可以直接转移到监控视频下进行行人再识别，并取得了很有竞争力的结果。最近，Li 等人[57] 收集了一个大型数据集，其中包含丰富的注释行人属性，以促进基于属性的行人再识别研究的发展。

深度方法 自 Krizhevsky 等人[10] 在 2012 年的 ILSVRC 比赛中取得了大幅领先的成绩以来，基于卷积神经网络的深度学习模型一直就很受研究者们的青睐。最早在行人再识别领域中使用深度学习的工作是文献[13,58]。一般来说，被广泛使用的卷积神经网络模型主要有两种，第一种类型是用于图像分类[10] 和物体检测[59] 的分类模型。以上提到的模型都是属于端到端（end-to-end）的特征学习模式，有一些其他的工作则在此基础上再加入了一些低级特征作为卷积神经网络的输入来辅助训练。2017 年，在文献[60]中，Wu 等人为每一张图像分别产生一个 Fisher 向量[61]，该向量由 SIFT 和颜色直方图这样低级的描述子组合而成。这种混合网络在输入的 Fisher 向量上构建全连接(Fully Connected，FC）层，并采用线性判别分析（LDA）作为目标函数，以产生具有低类内方差和高类间方差的特征表示。Wu 等人[62] 还提出将 FC 特征和低级特征向量

进行拼接，并在其与最终 softmax 层之间再加入一层全连接层，此方法使用传统手工特征来约束神经网络的 FC 特征。

1.2.2 基于度量学习的行人再识别方法

基于度量学习的行人再识别方法的主要目标在于学习一些能够对样本进行区分和匹配的距离度量函数，以及通过施加一些度量限制来加强特征提取器的特征提取的鉴别性和鲁棒性。该类方法同样可以根据是否使用深度卷积神经网络学习技术分为经典方法和深度方法，其具体的一些介绍如下：

经典方法 在基于手选特征的行人再识别系统中，良好的距离度量扮演着至关重要的角色，因为高维视觉特征通常会主要包含的是样本之间的差异而不会或很少会包含样本差异下的不变因子。文献[63]对度量学习方法的进行了很全面总结。这些度量学习方法可以根据监督信息的多少分为全监督学习（supervised）与无监督（unsupervised）学习，全局学习（global learning）与局部学习（local learning）等。在行人再识别任务中，大多数工作属于监督全局距离度量学习的范围。全局度量学习的一般思想是使得同一类的所有向量之间的距离尽量接近，同时将不同类的向量进一步分开。最常用的公式是马哈拉诺比斯距离（Mahalanobis distance）函数或其变型，该距离在欧氏距离（Euclidean distance）的基础上使用线性拉伸和特征空间旋转来对欧氏距离进行了推广。两个向量 x_i 和 x_j 之间的平方距离可写为:

$$d(x_i, x_j) = (x_i - x_j)^T \mathbf{M}(x_i - x_j) \tag{1.1}$$

其中 $\mathbf{M}$ 是一个正半定基。公式(1.1)可以表示为 Xing 等人[64] 提出的凸规划问题。

在行人再识别领域,目前最流行的度量学习方法,要数在 2012 年被提出的 KISSME[65]。KISSME 就是基于公式(1.1)。具体来讲，文献[65]将关于一对 $(i\ j)$ 是否相似的决定看成似然比检验。首先定义成对差为 $(x_{i,j} = x_i - x_j)$，然后假设差异空间是具有零均值的高斯分布。文献[65]中则表示，马哈拉诺比斯距离度量可以自然地从对数似然比检验得出，并且在实践中，通常还会采用主成分分析（PCA）来消除维度的相关性。

基于公式(1.1)还提出了许多其他度量学习方法。在早一些时候，一些经典的度量学习方法主要针对的是最近邻居分类任务。Weinberger 等人[66] 在 2009 年提出了大边际最近邻学习（large margin nearest neighbor，LMNN）方法，该方法为目标邻居（正确匹配）建立边界并惩罚侵入其中的那些错误匹配。该方法属于有监督的局部距离度量学习类别。为了避免 LMNN 中遇到的过度拟合问题，Davis 等人[67] 提出信息理论度量学习（information-theoretic metric learning，ITML）作为满足给定相似性约束和确保学习度量接近初始距离函数之间的权衡。

随之，Hirzer 等人[68] 提出放正性约束来为在小计算成本下的情况下得到矩阵 M 的近似。Chen 等人[47] 除了马哈拉诺比斯距离之外还增加了双线性相似性，因此可以对块间 (cross-patch) 相似性进行建模。在文献[39]中，Li 等人对全局距离度量与局部自适应阈值规则进行了耦合，其中局部自适应的阈值规则中包含了 (x_i, x_j) 的正交信息。在文

献[69]中，Liao 等人提出保留半正定约束，并建议对正样本和负样本进行不同的权值分配。Yang 等人[70] 则考虑图像对之间的差异性和共性，并表明不相似对的协方差矩阵可以从相似对的协方差矩阵推断得到，这使得学习过程可扩展到大数据集。

除了学习距离度量之外，一些工作专注于学习判别性子空间。Liao 等人[12] 提出学习一个投影将 w 投影到一个低维子空间，利用交叉视图数据并以可以用与线性判别分析（LDA)[71] 类似的方式进行求解，

$$\mathbf{J}(\mathbf{w}) = \frac{\mathbf{w}^T\mathbf{S}_b\mathbf{w}}{\mathbf{w}^T\mathbf{S}_w\mathbf{w}} \tag{1.2}$$

其中 $\mathbf{S}_b$ 和 $\mathbf{S}_w$ 分别是类间和类内散布矩阵。然后，通过使用 KISSME 在得到的子空间中进行距离函数的学习。为了学习 w，Zhang 等人[19] 进一步采用零 Foley-Sammon 变换 (null Foley-Sammon transform) 来学习一个判别零空间，其满足零内散射和一个正类间散布。为减少尺寸，Pedagadi 等人[48] 在 2013 年结合无监督 PCA（主成分分析）和监督局部 Fisher 判别分析，保留了局部邻域结构。在文献[36]中，Mignon 等人提出了成对约束分量分析（PCCA)，它学习的是一个线性映射函数，其能够直接处理高维数据，而 ITML 和 KISSME 应该先进行降维步骤。在文献[72]中，Xiong 等人进一步提出了两种现有子空间投影方法的改进版本，即正则化 PCCA[36] 和内核 LFDA[48]。除了基于 Mahalanobis 距离 (见公式(1.1)) 的方法，也有些使用其他学习工具，如支持向量机 (SVM) 或 boosting 算法。Prosser 等人[33] 提出来先学习一组弱的排序 SVM，随后组装成一个更强大的排序机制。在文献[73]中，Liu 等人提出使用一个结构性的 SVM 来对不同的颜色描述符进行组合。Zhang 等人[51] 在 2016 年通过对每个行人身份 ID 学习训练一个特定的 SVM，并将每个测试图像根据其视觉特征将其映射成一个权重向量。Gray 和 Tao[31] 提出使用 AdaBoost 算法选择并将许多不同类型的简单特征组合而成一个相似度函数。Zhu 等人[30] 考虑到不同困难程度的不匹配样本之中所包含的大量有效信息，通过利用一个对称的三元组约束来进行更好的距离学习。随后，Zhu 等人[74] 又提出了基于集合距离学习的行人视频再识别方法，该方法同时考虑了行人视频的类内差距以及类间差异，并取得了比较好的再识别效果。

深度方法 基于度量学习的深度学习方法大部分都是使用图像二元组[75] 或三元组[76] 作为输入的孪生网络模型（Siamese Network)。在一些大规模行人再识别数据库被提出来之前，大多数行人再识别数据集仅为每个行人提供有限数量的训练图像，例如 VIPeR[31]，因此当时大多数基于卷积神经网络的行人再识别方法都采用的是孪生网络结构。在文献[58]中，每幅输入的图像被划分为三个互相有重叠的水平图像块，并且这些部分经过两个卷积层加上一个全连接层，最后被融合在一起并输出一个该图像的特征表示。最终使用余弦距离来计算两个特征向量之间的相似度。2014 年，Li 等人[13] 设计的架构的不同之处在于其添加了一个块匹配层，它将两个不同图像的水平条纹的卷积响应相乘，类似于 ACS[38]。后来，Ahmed 等人[77] 通过计算交叉输入邻域差异特征来改进孪生网络模型，该方法将来自一个输入图像的特征与另一个图像的相邻位置中的特征

进行比较。文献[13]使用卷积乘法来计算相似高度的图像块相似性，而 Ahmed 等人[77]则使用的是减法。Wu 等人[78] 使用较小尺寸的卷积滤波器使得网络层数得以变深，该网络被称为 “PersonNet”。在文献[79]中，Varior 等人将长短期记忆（Long-short Term Memory ,LSTM）与孪生网络结合在一起，该方法首先利用 LSTM 来顺序处理图像块，以便可以记忆存储空间信息以增强深度特征的鉴别能力。Varior 等人[80] 则提出在每个卷积层之后插入门控函数（gating function），使得在输入一对测试图像的时能够捕获到细微的有效特征。该方法好几个数据集上都达到了最高的准确率，但其缺点也很明显。查询图片必须在输入到网络之前与图库中的每一张图像进行配对，这在大型数据集上是一个很花费时间并且效率低下的过程。类似于文献[80]，Liu 等人[81] 提出在孪生网络中集成基于软注意的模型，使其具有能够自适应地关注输入图像对的重要局部部分的功能；然而，这种方法也会导致计算效率的低下。上述这些工作都要求输入都是图像二元组，在 2016 年，Cheng 等人[82] 设计了三元组损失函数，将三张图像作为输入。在第一卷积层之后，对于每个图像划分四个重叠的身体部分，并且在 FC 层中与全局的一部分融合。Su 等人[83] 提出了一个三段学习过程，其中包括使用独立数据集来进行的属性预测和在具有 ID 标签的数据集上训练的属性三元组损失。

针对二元组输入的孪生网络，其损失函数可以写成如下形式：

$$L(W,(y,x_i,x_j))=\frac{1}{2N}\sum_{n=1}^{N}D_W^2+(1-y)\max(m-D_W,0)^2 \tag{1.3}$$

其中 $D_W=\parallel X_1-X_2\parallel_2=(\sum_n^N(x_i^n-x_j^n)^2)^{\frac{1}{2}}$ 代表两个样本特征 x_i 和 x_j 的欧氏距离，y 为两个样本是否匹配的标签，$y=1$ 代表两个样本相似或者匹配，$y=0$ 则代表不匹配，m 为设定的边际阈值。这种损失函数最初来源于 Hadsell 等人[84]，主要是用在降维中，即本来相似的样本，在经过降维（特征提取）后，在特征空间中，两个样本仍旧相似；而原本不相似的样本，在经过降维后，在特征空间中，两个样本仍旧不相似。

通过观察公式(1.3)可以发现，这种损失函数可以很好的表达成对样本的匹配程度，也能够很好地用于训练提取特征的模型。

与之相对应的，Schroff 等人[76] 在人脸识别任务上提出了一种基于三元组的损失函数，其可以写成以下形式：

$$L(W,(x_i,x_j,x_k))=\sum_i^N[\parallel f(x_i^n)-f(x_j^n)\parallel_2^2-\parallel f(x_i^n)-f(x_k^n)\parallel_2^2+\alpha]_+ \tag{1.4}$$

其中要求输入 x_i 和 x_j 是属于同类的样本，而 x_k 是属于与 x_i,x_j 不同类的样本。一般来说，x_i 被称为锚 (Anchor) 样本，x_j 被称为正 (positive) 样本而 x_k 被称为负 (Negative) 样本。这样设计的初衷就在于希望模型能够让锚样本在特征空间中和正样本之间的距离比与负样本之间的距离要来的小，并且需要满足一个边际值 α，即：$\parallel f(x_i^n)-f(x_j^n)\parallel_2^2<\parallel f(x_i^n)-f(x_k^n)\parallel_2^2-\alpha$，该式中的 α 则是一个边际阈值 (margin)，它的作用是限制两对正负样本之间的分离程度。通过优化公式(1.4)，可以使得学习出来的特征具有以下两个特

点：1）不同类之间的样本特征距离拉得很开，也就是不同类别样本更不容易混淆；2）同类之间的样本特征距离靠得比较紧密。这样的目标函数设计对于模型的鲁棒性的提高是比较有帮助的。

尽管如此，孪生网络模型的缺点是它并没有充分利用到行人再识别的行人标注。事实上，孪生网络模型只考虑到了图像二元组（或三元组）的标签。仅仅知道图像对是否相似（属于相同的身份）是只是用到了行人标签中的一部分弱信息。另一种可能有效的策略就是同时使用分类和验证模式，从而达到能够充分利用行人再识别标签的目的。在文献[85]中，Xiao 等利用多个数据集的训练集组成一个全新训练集，并且在分类网络中采用 softmax 损失函数。结合为每个全连接神经元生成的影响分数并根据该影响分数的基于区域的随机失活（dropout），所学习到的特征表示得到了很高的行人再识别准确率。基于这样的考虑，文献[14]和文献[86]均表明分类模型在更大的数据集上，例如 PRW 和 MARS，在不需要进行精细的训练样本筛选的情况下就可以取得良好的性能。然而，使用分类损失函数进行训练就需要能为每个行人能够提供更多的训练样本来促使模型收敛。除此之外，针对于度量学习的方法还具有一个缺陷就在于其图像元组的构建，对于大型的行人再识别数据库来说，行人元组的组合数量是一个呈指数型的增长趋势，这就不可避免地带来了巨大的数据量训练的问题，在这样一个巨大数据量的训练情况中还包含正负样本对数量不均等等问题需要在模型设计时加入考虑。

1.2.3 再排序 (re-ranking)

再排序 (Re-ranking) 作为一种提高物体检索准确性的方法已经得到了充分的研究[11,87,88]。其中，许多工作利用 k-最近邻来探索检索样本的相似性关系以解决再排序问题。Chum 等人[89] 在 2007 年提出了平均查询扩展（Average query expansion, AQE）方法，其中通过对前 k 个返回结果中的向量求平均来组成一个新的查询向量，并用其来重新查询数据库，该方法具有更强的鲁棒性。此外，为了利用那些与查询样本距离较远的不匹配样本，Arandjelovic 和 Zisserman[90] 提出了判别式查询扩展（discriminative query expansion, DQE）以使用线性 SVM 来获得权重向量，并且根据到决策边界的距离来对初始排序列表进行修改。Shen 等人[91] 利用原始检索排名列表的 k-最近邻作为新查询来产生新的排名列表，并根据其在所生成的排名列表中的位置来计算每个图像的新的相似度评分。最近，稀疏上下文激活（Sparse Contexual Activation，SCA）[92] 提出了用矢量的形式来表达近邻集，并通过广义 Jaccard 距离来度量样本相似性。为了防止在前 k 的检索列表中的错误匹配对最终列表造成不好的影响，在文献[93,94]中采用了 k-互惠 (reciprocal) 最近邻的概念。Jegou 等人[93] 提出了上下文相异性度量（Contextual Dissimilarity Measure，CDM），并通过迭代地归一化每个点到其近邻的平均距离来对它们的相似性进行修正。Qin 等人[94] 正式提出了 k-互惠（reciprocal）邻近的概念。k-互惠邻近被认为是与待查询样本具有高度的相似性的样本，它们可以用于构造一个闭集并对剩余的数据集进行重新排序。文献[95]的工作在这两个方面来说与前者不大相同。其并不利用对称的最近邻域关系来修正相似性[93]，或者直接将 k-互惠最近邻视为排名靠

前的样本，与文献[94]相反，文献[95]通过比较它们的 k-互惠最近邻居来计算两个图像之间的新距离。

最近，一些研究工作[96–105] 已经注意到基于再排序的方法在行人再识别领域中的应用。文献[99,106]和文献[107]的再排序方法需要人工交互或标签监督，而文献[95]提出的则是主要是专注于自动和无监督的解决方案。Li 等人[101] 通过分析每个图像对其分别的近邻的相对信息和直接信息来构建重新排序模型。文献[102]提出了通过联合考虑排名列表中的内容和上下文信息来学习无监督的重新排序模型，这有效地去除了模糊样本以提高行人再识别的性能。Leng 等人[103] 提出了一种双向排序方法来融合内容和上下文相似性来计算新相似度，该相似度被用来对初始排名列表进行修正。最近，文献[104,105]中提出了用共同最近邻居不同基准方法来研究行人再识别任务。Ye 等人[104] 将全局和局部特征的共同最近邻进行组合作为新的查询样本，并通过聚合全局和局部特征的新排名列表来修正初始排名列表。在文献[105]中，作者利用最近邻集来计算不同基准方法的相似性和不相似性，然后进行相似性和不相似性的聚合以优化初始排序列表。然而，使用 k 最近邻居直接实现再排序可能会影响到整体的性能，因为最近邻中通常包括很多的错误匹配。这些提到的方法在再排序方面已经取得了很有潜力的进展，同时也为从最近邻居发现更多信息做出了贡献。

表 1.2 现有的图像行人再识别数据库

数据集	时间	ID 个数	图片张数	相机个数	标定方式	评价
VIPeR[108]	2007	632	1264	2	手工	CMC
iLIDS[17]	2009	119	476	2	手工	CMC
GRID[109]	2009	250	1275	8	手工	CMC
CAVIAR[110]	2011	72	610	2	手工	CMC
PRID2011[111]	2011	200	1134	2	手工	CMC
WARD[112]	2012	70	4786	3	手工	CMC
CUHK01[113]	2012	971	3884	2	手工	CMC
CUHK02[114]	2013	1816	7264	10(5 对)	手工	CMC
CUHK03[13]	2014	1467	13164	2	手工/DPM	CMC/mAP
RAiD[44]	2014	43	1264	4	手工	CMC
PRID 450S[115]	2014	450	900	2	手工	CMC
Market-1501[8]	2015	1501	32668	6	手工/DPM	CMC/mAP
DukeMTMC-reID[116]	2017	1404	36411	8	手工	CMC/mAP

1.3 行人再识别数据集和评价标准

目前关于行人再识别的任务都在一些公开的标准数据集上进行试验，并且通过一系列的评价标准来评估所提出的方法的性能。在下面两个小节中，分别介绍行人再识别的

公开数据集和广泛采用的评价标准。

1.3.1 行人再识别数据集

在行人再识别领域同样也出现了很多的公共数据集，这些数据集中既有基于图像的也有基于视频的。这些行人再识别数据集分别在表1.2和表1.3中做了展示。在之前被广泛使用的是 VIPeR[108]。VIPeR 数据集总包含了 632 个行人 ID，每个行人 ID 包含 2 张不同摄像头拍摄的图片。训练测试集由 10 次随机的数据分割构成，训练集和测试集都分别有 316 个 ID，随机的训练测试集分割可以检测方法的泛化能力。此外，针对不同的使用场景也有不同的数据集被提出来。例如，GRID[109] 数据集是在地铁站采集的，iLIDS[17] 是在机场到达大厅采集的，CUHK01[113]，CUHK02[114]，CUHK03[13]，Market-1501[8] 和 DukeMTMC-reID[116] 分别是在香港中文大学，清华大学和杜克大学的校园内被采集的。基于视频的 MAR[14] 和 DukeMTMC-VideoreID[25] 数据集则分别是在 Market-1501[8] 和 DukeMTMC-reID[116] 的扩展，是在其基础之上构建的。

目前在行人再识别领域被广泛采用的大规模公开图像数据集主要包括了 CUHK03[13]，Market-1501[8]，DukeMTMC-reID[116]，其原因在于行人身份多，图片数量多。因此，本文的主要实验均在上述的图像数据集上进行实验，为了节省每章的数据集介绍部分，我们将具体介绍一些大规模的行人再识别数据库。而视频再识别数据库本身是图像再识别数据库的扩展，也仅仅无监督的情况下使用到，因此在这里并不进行详细介绍，其具体比较将在第4章中进行。下面介绍的是三个基于图像的行人再识别数据库。

表 1.3 现有的视频行人再识别数据库

数据集	时间	ID 个数	视频个数	相机个数	标定方式	评价
ETHZ[117]	2007	148	148	1	手工	CMC
3DPES[118]	2011	200	1,000	8	手工	CMC
PRID2011[111]	2011	200	400	2	手工	CMC
iLIDS-VID[119]	2014	300	600	2	手工	CMC
MARS[14]	2016	1261	20715	6	DPM/GMMCP	CMC/mAP
DukeMTMC-VideoReID[25]	2018	702	4832	8	手工	CMC

CUHK03 行人再识别数据库[13] CUHK03 数据库的提出是因应深度学习网络对训练数据量的需求，因为在此之前已有的数据集都太小，无法训练深度神经网络。CUHK03 数据集中包括了 1360 个行人的 13164 张图像。除了手动裁剪的行人标注框外，CUHK03 还提供了基于 DPM[120] 行人检测算法检测到的标注框样本。在这个数据集中，标注框未对齐、遮挡和身体部位缺失是很常见的。在 CUHK03 数据集中，从多对摄像机视图中收集的样本都是混合的，它们形成复杂的交叉视图变换。此外，即使是在单个摄像机视图内，也有光照变化，这些可能是由天气，太阳方向和阴影分布引起的。

CUHK03 的测试协议有两种。第一种为旧的版本[13]，此测试协议为数据集的默认设置。具体地说，即随机选出 100 个行人作为测试集，1160 个行人作为训练集，100 个

表 1.4 CUHK03 数据集

	总数据集	CUHK03 Labeled			CUHK03 Detected		
		训练集	查询集	候选集	训练集	查询集	候选集
旧数据集分割标准							
行人个数	1467	1160	100	100	1160	100	100
图片张数	14097	-	-	-	-	-	-
新数据集分割标准							
行人个数	1467	767	700	700	767	700	700
图片张数	14097	7368	1400	5328	7365	1400	5332

行人作为验证集（这里总共 1360 个行人而不是 1467 个，这是因为实验中没有用到摄像头组 4 和 5 的数据），重复实验二十次，取所有实验结果的平均值作为最终的结果以消除实验分割随机性对实验结果的影响。第二种测试协议[95] 类似于 Market-1501，它将数据集分为包含 767 个行人的训练集和包含 700 个行人的测试集。在测试阶段，随机选择一张图像作为查询图像，剩下的作为候选集，这也就可以将设定与 Marke-1501 统一起来。

Market-1501 行人再识别数据库[8] Market-1501 数据集是在清华大学校园的一

表 1.5 Market-1501 数据集

	Market-1501 数据集			
	总数据集	训练集	查询集	候选集
行人个数	1501	751	750	750
图片张数	32688	12936	3368	16384

个超市收集的，收集过程中一共使用了六台摄像机，放置在校园超市的前面。该数据集包含 1,501 个行人的 32,668 个标注框。数据集中的行人都至少出现在了两个摄像头中，以便满足行人再识别的实验设定。相较于之前使用裁剪的标注框的数据集来说，Market-1501 数据集采用的是 DPM[120]，其原因在于模拟现实应用中行人检测的标定框也会存在的偏移与不对齐的情况。Market-1501 数据集的整体数据统计可以参照表1.5。

DukeMTMC-reID 行人再识别数据库[116] DukeMTMC-reID 是基于一个多目标，多摄像机行人跟踪数据集 DukeMTMC[121] 构建而成的，其是 DukeMTMC 的子集。原始的视频数据集包含来自八个不同相机的八个 85 分钟高分辨率视频。这些视频都提供了手动标定的行人标注框。该行人再识别数据库是通过提取视频中的图像来组成的一其组成方法也参照采用了 Market-1501 数据集[8] 的格式。行人图像的采集具体标准为

表 1.6 DukeMTMC-reID 数据集

	DukeMTMC-reID 数据集			
	总数据集	训练集	查询集	候选集
行人个数	1404	702	702	702
图片张数	36411	16522	2228	17661

每隔 120 帧从视频中进行裁剪，一共生成了 36,411 个行人标注框，它们的 ID 沿用文献[121]中的。因此，DukeMTMC-reID 数据集具有来自八个摄像机的 1,812 个行人。在两个以上的摄像机中出现的行人身份（ID）一共有 1,404 名，仅仅在一个摄像机中出现的有 408 名。接着，随机选择 702 个行人 ID 作为训练集，剩下的 702 个行人 ID 作为测试集。在测试集中，从每个摄像机中的每个 ID 选择一个查询图像，并将剩余的图像放入待候选集中。结果，该数据集总共有 702 名行人的 16,522 张训练图像，702 名行人的 2,228 张查询图像和 17,661 张候选集图像。

上述三个常用大型行人图像再识别数据库的示例图片可参照图1.4。

(a) CUHK03 labeled 数据集 (b) CUHK03 detected 数据集

(c) Market-1501 数据集 (d) DukeMTMC-reID 数据集

图 1.4 常用数据集样本示例

1.3.2 行人再识别评价指标

在对行人再识别算法的性能进行评价时，通常使用的是累积匹配特性（Cumulative Match Characteristic, CMC）曲线。曲线综合反映了分类器的性能，它的评价指标与现在深度学习分类问题中中常用的 Top1 错误率或 Top5 错误率的评价指标具有相同的意思，但不同的是在于 Rank-1 识别率表示的是正确率而不是错误率，两者的关系是 Rank1 识别率 =1-Top1 错误率，Rank5 识别率 =1-Top5 错误率。也就是说，CMC 表

示的是匹配目标出现在不同指定大小的候选列表中的概率。无论检索集中有多少正确的匹配，CMC 计算中只计算第一个匹配项。所以基本上，只有当每个查询存在一个正确匹配候选时，CMC 才能作为评估方法准确。在实践中，当人们更倾向于希望在排序匹配列表的顶部位置得到真实的匹配。在这里，CMC 通常也可以用另外一种形式更加直观的给出，那就是 rank-k 准确率，也就是说正确匹配出现在匹配列表的前 k 位的概率。本文中亦都使用 rank-k 准确率来评估行行人再识别算法的性能。

然而，对于性能评估的完整性，当检索集中存在多个相应匹配时，Zheng 等人[8] 则建议使用平均精度（mAP）来进行评估。其动机是在于一个完善的行人再识别系统应该能够将所有真正的匹配都检索出来给用户。所以可能存在的情况是：两个算法对于发现第一个匹配具有相同的能力，而它们之间具有不同的检索召回（Recall）能力。在这种情况下，CMC 或 rank-k 准确率就会没有足够的判别能力，但 mAP 却可以将其优劣区分开来。因此，mAP 与 rank-k 准确率结合在一起作为行人再识别的评断标准，能够更加全面的对算法的性能进行评价。在此之后，这两种性能评价指标被后续的工作[79,80,122] 所采用，并渐渐成为普遍适用的评价指标。

1.4 本文的主要研究工作

在前面几节中，本文对行人再识别领域的一些发展起源和发展现状存在的一些挑战进行了简单的回顾和分析。虽然，现有的行人再识别方法可以有效地提取行人数据的特征信息，并且也取得了一些比较好的效果，但是这些特征提取方法都面临一些鲁棒性以及数据依赖性的问题。具体地讲，现有的行人再识别方法具有以下限制：

(1) 经典的基于数据库的全监督行人再识别算法均是直接采用分类模型进行样本特征的提取，然后进行行人图像的检索。这样的方法虽然有着不错的性能表现，但由于单支的网络的表达能力有限（更加偏重于输入图片的全局特征），而行人再识别领域属于细粒度的图像分类问题，其对于局部的特征要求比较高，所以其特征的完备性以及鲁棒性并不是最优的。因此，如何学习到一个更加全面能够兼顾全局和局部的特征表示显得尤为重要。

(2) 深度卷积网络的训练学习都是由数据驱动的，足够量的训练数据才能够保证模型的参数学习具有更强的泛化性。通用的解决方案有数据增强。生成对抗网络作为一种能够生成新样本的方法也受到了广泛的关注。但是，目前对于如何使用由生成对抗网络产生的合成图片作为数据增强方法的研究还刚刚起步，目前的方法并不能很好的利用到合成图片的一些特性。因此，如何使用一种高效的方法对这些生成图片进行合理的利用也是一个非常重要的研究问题。

(3) 有效的无监督特征学习能够摆脱对于标注数据的依赖性。而在行人再识别领域，标注数据的获取往往具有很高的代价，需要大量的人力，物力和财力的投入。现有的无监督行人再识别的很多尝试还是基于额外的带标注的行人再识别数据库上进行的，只有极少数的研究是完全无监督的设定。因此，能够进行完全无监督的

行人在识别也一直被研究者们所关注着。

本文基于不同的监督情况下的行人再识别任务，对其特征学习与表示以及网络结构设计的相关理论和算法进行了研究和探索，将不同方面的设计相联系，具体地讲，本文所作的工作大致可以归纳为以下几个方面:

(1) 在全监督的行人再识别设定下，提出了一种新的网络结构：该网络结构是一种类孪生网络的结构，通过它能够使得网络学习到互补的特征。具体来讲，传统单支网络的特征学习主要集中在对图片全局最具鉴别特征的区域进行学习，而提出的该网络结构包含两个分支，分别对不同粒度的行人特征进行学习与建模。主干网络的主要功能和传统单支网络一样是专注在学习全局的特征信息，而次干网络则通过在主干网络的指导学习下对图片中所包含的局部信息进行挖掘。为了进一步地保证两支网络特征的差异性，又在两支结构的基础上加上了成对排序误差。

(2) 在半监督的行人再识别设定下，提出了一种新的基于特征相似性的伪标签生成方法。具体来讲，传统的伪标签生成方法大多使用的是类别概率，该方法的缺点在于概率仅仅衡量属于某一类的可能性，并没有考虑到在特征空间中的相似性，此外，该基于概率的伪标签方法无法直接生成分布式的伪标签，具有相当的局限性。基于特征相似性的伪标签生成方法首先提出了一个多任务损失函数，能够联合考虑特征空间中的类内和类间的变化，并同时生成伪标签。此外，基于特征相似性的计算方法使得两种不同的标签编码模式在同一个框架下生成。

(3) 在无监督的行人再识别设定下，提出了一种基于分散度的凝聚聚类方法来挖掘潜在的数据分布特点从而能够对无标签数据进行利用，最终得到鉴别鲁棒的特征提取模型。具体来讲，在聚类的过程中，簇融合的选择判断标准是基于两种考虑，尽量小的簇间分散度和尽量小的簇内分散度。该选择标准优于其他的聚类选择标准地方在于其是一个更加全局的考虑，并且具有单独点优先聚类和防止坏簇形成的优点。利用逐步的聚类能够逐步地发现数据中潜在的分布，从而逐步地提升无监督的行人再识别的性能。

1.5 本文的章节安排

本文围绕不同数据监督情况下基于深度学习的行人再识别的关键技术展开研究，主要从有监督的互补特征提取、半监督的伪标签生成和无标签的聚类等几个不同的方向进行了研究与探讨。本文的内容结果关系如图1.5所示，其主要可以分为三大部分，分别是绪论、主要研究内容和总结。绪论部分主要回顾了行人再识别的发展历程以及对目前该领域的研究的进展做出了综述。主要内容研究包含了针对不同的数据监监督前提下存在的不同的亟待解决的问题分别提出的创新性工作，并按照监督程度的由重到轻进行了排序，即从全监督到半监督，最终再转至无监督。最后一部分则对文章进行了总结并给出了后续的一些工作方向。

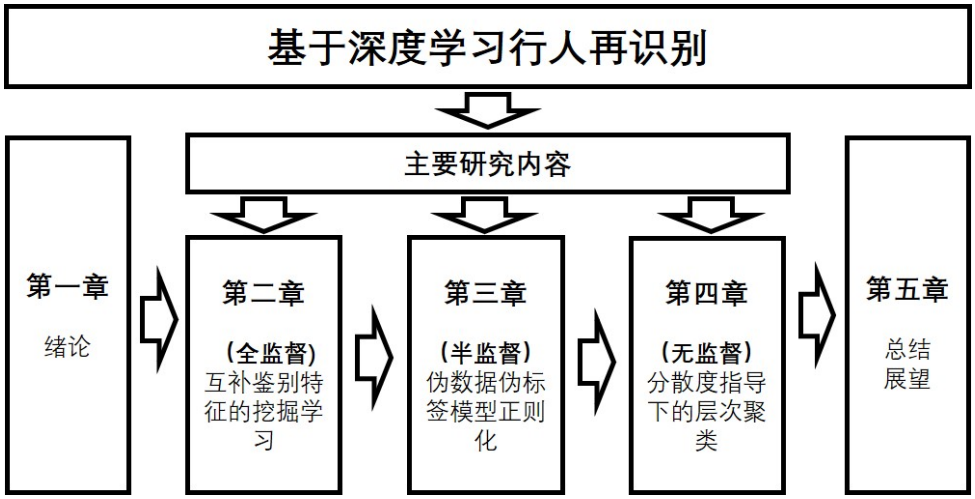


图 1.5 文章结构关系图

具体来讲，后续章节内容安排如下：

第 2 章 互补鉴别特征提取的全监督行人再识别：针对单支网络结构提取的特征存在着鉴别信息不足的缺点，提出了一种互补鉴别特征提取的算法。具体来说，在孪生网络的基础上提出了能够促使网络进行互补鉴别信息挖掘的特征掩码模块，通过从主干网络的高层特征表示映射出分支网络的特征选择权重从而学习不同鉴别特征；在此基础上，又提出在结构上加上成对排序误差，使得最终所得鉴别特征更加具有互补性。最后，结合实验结果和可视化直观地验证了该方法的有效性。

第 3 章 伪数据伪标签正则化深度模型的半监督行人再识别：针对现有行人再识别数据库行人标注数据少的缺点，提出了利用伪数据伪标签正则化模型的算法。该章首先介绍了利用生成对抗网络进行行人伪数据的生成，然后在此基础上，针对如何为生成的伪数据生成伪标签使其能够参与训练，提出了一个通用伪标签生成框架。具体来讲，提出了利用特征空间中的无标签样本与有标签样本之间的特征相似度作为伪标签生成的依据，得益于特征相似性的度量，该标签生成方法可以在一个统一的框架下生成两种不同伪标签编码形式。所提出的方法在相关行人再识别数据集上的实验得出了比类似方法高的性能，从而证明了该方法的有效性。

第 4 章 基于分散度的无监督行人再识别：针对现有的行人再识别数据库数量少(对比于实际应用)，标注难的问题，提出了一种基于分散度的凝聚聚类无监督算法。具体来说，利用簇间分散度衡量簇间的相似性，簇内分散度作为簇内紧凑度的正则项，来对聚类候选簇进行挑选。该方法不仅能够优先选择单独的数据点，还能够阻止坏簇的形成与蔓延。除此之外，该方法还与卷积神经网络有着相互促进的作用，提升了网络的学习速度以及鲁棒性。最后，在基于图像和视频的数据集上进行了实验，证实了该方法的有效性。

第 5 章 总结了全文工作，对本文研究的不足之处做了评述；简单分析了文中一些未展开的相关工作，并对今后的研究工作做了展望。

2 基于互补鉴别特征提取的行人再识别

与众多图像模式识别系统类似，特征提取器是传统图像处理的一个重要模块。一个好的特征提取器可以计算出较好的图像特征，这些特征相比原始像素数据具有更丰富的鉴别信息。如今，在行人再识别领域已经出现了很多基于深度学习的特征提取器，可是很多特征提取器仅仅是从分类任务直接迁移过来的。而在行人再识别任务中，测试集的行人与训练集的行人并无重叠，因此也对特征提取器的鲁棒性提出了更好地要求。由此可见，在行人再识别领域，如何提取出具有丰富鉴别信息的行人特征表示是一个非常关键的问题。本章基于有监督情况下的鲁棒特征的学习，提出了一种基于特征掩码的网络方法（Feature Mask Network, FMN），从而实现了互补鉴别特征的提取。

2.1 引言

现有的深度学习网络能够在分类任务上取得不错的性能，表明了其提取的特征具有一定的鉴别性。尽管如此，在 2017 年，Singh 等人[123] 指出现有的深度学习网络倾向于聚焦于图片中最具鉴别性的部分，例如在识别包含狗的图片的时候，网络的注意力 (Attention) 主要集中在狗的面部。相似地，该类方法应用在行人再识别问题上也会有同样的问题，学习到的网络更多的注意力集中在了行人的躯干部分。然而，行人再识别与普通分类问题数据分布上有着极大的差异性，即：行人再识别所有图像内容均为行人，他们之间仅仅有着较小的差距，而普通数据分类涵盖了各种各样内容差异巨大的自然图片。所以，作为一个鲁棒的行人再识别的特征提取器需要具有能够挖掘细粒度（fine-grained）特征的能力，如何能够尽可能多地进行鉴别特征的提取就成了一个行人在识别领域很重要的研究问题。

此外现有的普遍的行人再识别方法假设行人标定框是由专用检测器检测出来的。然而，这样的检测并不总是完美的，导致诸如在标注框中包含过多背景，行人身体的不完全覆盖和局部化不匹配的问题。在监视场景中存在部分遮蔽行人的重度遮挡这一事实更是加剧了这种情况。本章所提出的工作的动机是如何挖掘能够克服这些挑战的关键信息，即同时关注重要但可能微妙的局部细节和全局特征。在本章中，将对一种自动的特征掩码方法进行描述，该方法着重于使用深度神经网络学习全局的图像特征表示以及互补的局部细节特征。这样的识别算法能够挖掘更多的局部区域，而这些局部区域包含更多对身份预测任务或特征提取的更有价值和辨别力的线索。本章提出了一种基于特征选择的软注意力模型 (soft attention) 策略在一个端到端的学习框架下进行了不同粒度的特征提取。除了避免由于不完美的行人标注框引起的问题之外，本章所提出的网络还通过上述软注意力模型将注意力转移到各个局部区域来学习到更有助于鉴别性的特征（例如两个不同行人穿着类似的服装，但是穿不同颜色的鞋子）。因此，该网络结构能够利

用已经学习的全局判别特征作为指导和利用特征掩码网络的动态选择机制，通过给低层次的特征赋予不同的权重来实现更紧凑的人类特征表示。

本章聚焦于有监督情况下行人更加鉴别鲁棒的特征提取，综合了细粒度行人再识别和神经网络集成的思想，提出了一种互补鉴别特征提取的新的网络结构，与之前的工作相比较的话，本章所提出的方法具有以下优势：

(1) 本章提出了一种特征掩码网络（FMN），它可以动态地加强图像中的局部细节的挖掘，并提取出与全局特征相互互补的特征，两者组合在一起作为最终行人特征表示可以用来改善行人再识别的性能。
(2) 本章提出了一个多任务损失函数，它同时兼顾了优化分类以及基于单个输入的分支间成对排序损失，该排序损失可以确保网络特征的差异性，以得到高度鲁棒的特征描述。
(3) 本章所提出的方法与其他类似方法相比，在保证良好的性能的前提下，还具有易于实现和训练等优点。

2.2 通用特征提取深度模型

本节首先介绍一般通用的行人再识别深度特征提取模型以及行人再识别过程。具体来讲，一般基于卷积神经网络的行人再识别的可以分为两大部分，分别是行人鉴别分类网络的训练和基于行人特征的行人检索。在 2015 年，Zheng 等人[8] 提出，将行人再识别作为一种特殊的图像检索任务处理，目前这种模式已经被广泛采用。将查询集（probe）中的行人图像在候选集（gallery）中做检索，即：首先利用特征提取器对包含查询集和候选集中所有的行人图像进行特征提取并表示出来，然后计算查询样本与候选集中所有样本之间一对一的距离，再根据所得到的距离由小到大进行排序，最终返回排序的列表可以用来评估行人再识别方法的准确率。

行人鉴别分类网络的训练 近年来，卷积神经网络在图像分类任务中取得了不错的性能表现[10]。其一种典型的卷积神经网络结构如图2.1所示。一般卷积神经网络由输入层、卷积层、全连接层和损失函数层组成。输入层一般的输入是一个 $h \times w \times c$ 的图像，其中 h，w，c 分别是输入图像的高度宽度以及通道数（一般 RGB 图像的通道数为 3）。卷积层的功能是对输入数据进行特征提取，其内部包含多个卷积核，组成卷积核的每个元素都对应一个权重系数和一个偏差量。卷积神经网络中的全连接层等价于传统前馈神经网络中的隐含层。全连接层通常搭建在卷积神经网络隐含层的最后部分，并只向其它全连接层传递信号。特征图在全连接层中会失去 3 维结构，被展开为向量并通过激活函数（Activation Function）传递至下一层。在分类网络中，损失函数的选择一般是交叉熵损失函数。给定两个概率分布 p 和 q，通过 q 来表示 p 的交叉熵为：

$$H(p,q) = -\sum_{x} p(x) \log q(x) \tag{2.1}$$

其可以看成是两个概率分布之间的距离，或者说它刻画的是通过概率分布 q 来表达概

率分布 p 的困难程度，p 代表样本真实标签分布而 q 则代表的是预测值，交叉熵越小，两个概率分布越相近。

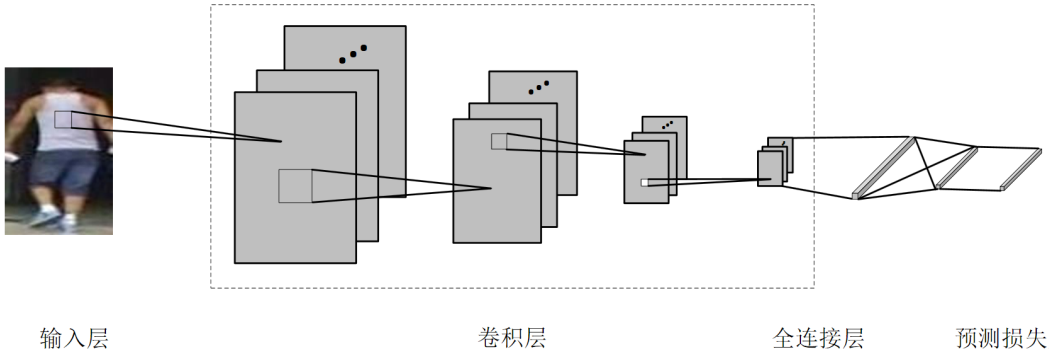


图 2.1 卷积神经网络结构示意图

网络结构选择 ResNet[124] 由 He 等人在 2015 年提出，在 ImageNet 比赛的图像分类任务上获得第一名，因为它“简单与实用”并存，之后很多方法都建立在 ResNet 的基础上。ResNet 卷积神经网络的结构在表2.1中被总结出来。基于行人再识别数据库中图片数据量的考量，ResNet-50 被广泛应用做基础网络。在本文中的所有实验的训练网络也都是基于 ResNet-50 的基础之上进行的。

表 2.1 ResNet 网络结构表

Layer name	output size	18-layer	34-layer	50-layer	101-layer	152-layer
conv1	112×112	7×7, 64, stride 2				
pool1	56×56	3×3 max pool, stride 2				
conv2_x	56×56	$\begin{bmatrix} 3 \times 3, 64 \\ 3 \times 3, 64 \end{bmatrix} \times 2$	$\begin{bmatrix} 3 \times 3, 64 \\ 3 \times 3, 64 \end{bmatrix} \times 3$	$\begin{bmatrix} 1 \times 1, 64 \\ 3 \times 3, 64 \\ 1 \times 1, 256 \end{bmatrix} \times 3$	$\begin{bmatrix} 1 \times 1, 64 \\ 3 \times 3, 64 \\ 1 \times 1, 256 \end{bmatrix} \times 3$	$\begin{bmatrix} 1 \times 1, 64 \\ 3 \times 3, 64 \\ 1 \times 1, 256 \end{bmatrix} \times 3$
conv3_x	28×28	$\begin{bmatrix} 3 \times 3, 128 \\ 3 \times 3, 128 \end{bmatrix} \times 2$	$\begin{bmatrix} 3 \times 3, 128 \\ 3 \times 3, 128 \end{bmatrix} \times 4$	$\begin{bmatrix} 1 \times 1, 128 \\ 3 \times 3, 128 \\ 1 \times 1, 512 \end{bmatrix} \times 4$	$\begin{bmatrix} 1 \times 1, 128 \\ 3 \times 3, 128 \\ 1 \times 1, 512 \end{bmatrix} \times 4$	$\begin{bmatrix} 1 \times 1, 128 \\ 3 \times 3, 128 \\ 1 \times 1, 512 \end{bmatrix} \times 4$
conv4_x	14×14	$\begin{bmatrix} 3 \times 3, 256 \\ 3 \times 3, 256 \end{bmatrix} \times 2$	$\begin{bmatrix} 3 \times 3, 256 \\ 3 \times 3, 256 \end{bmatrix} \times 6$	$\begin{bmatrix} 1 \times 1, 256 \\ 3 \times 3, 256 \\ 1 \times 1, 1024 \end{bmatrix} \times 6$	$\begin{bmatrix} 1 \times 1, 256 \\ 3 \times 3, 256 \\ 1 \times 1, 1024 \end{bmatrix} \times 23$	$\begin{bmatrix} 1 \times 1, 256 \\ 3 \times 3, 256 \\ 1 \times 1, 1024 \end{bmatrix} \times 36$
conv5_x	7×7	$\begin{bmatrix} 3 \times 3, 512 \\ 3 \times 3, 512 \end{bmatrix} \times 2$	$\begin{bmatrix} 3 \times 3, 512 \\ 3 \times 3, 512 \end{bmatrix} \times 3$	$\begin{bmatrix} 1 \times 1, 512 \\ 3 \times 3, 512 \\ 1 \times 1, 2048 \end{bmatrix} \times 3$	$\begin{bmatrix} 1 \times 1, 512 \\ 3 \times 3, 512 \\ 1 \times 1, 2048 \end{bmatrix} \times 3$	$\begin{bmatrix} 1 \times 1, 512 \\ 3 \times 3, 512 \\ 1 \times 1, 2048 \end{bmatrix} \times 3$
pool5	1×1	average pool, 1000-d fc, softmax				

特征提取过程和行人再识别过程 卷积神经网络经过上述的分类损失函数可以在行人再识别数据集上训练得到一组可以很好地进行行人分类的参数。基于这些参数，该网络可以对所有查询集和候选集中所有的样本进行特征提取，在 ResNet-50 上，每个行人特征都可以最终表达为一个 2048 维的向量。在行人特征提取完成之后，进行行人再识别的过程就是对于任意一个查询样本，根据其特征之间的距离对候选集中的样本进行

排序，最终返回行人再识别的一个列表。行人再识别的性能结果就可以通过对这个列表中的排序匹配情况进行统计而得到。

2.3 神经网络集成

神经网络已经在很多领域得到了成功的应用，但由于缺乏严密理论体系的指导，其应用效果完全取决于使用者的经验。1990 年，Hansen 等人[125] 提出了神经网络集成 (Neural Network Ensemble) 方法，并证明可以简单地通过训练多个神经网络并将其结果进行合成，显著地提高神经网络系统的泛化能力。1996 年, Sollich 等人[126] 为神经网络集成下了一个定义，即 “使用有限个神经网络对同一个问题进行学习，在某输入示例下的输出由各神经网络在该示例下的输出共同决定”。与仅创建一个模型[127] 相比，神经网络集合是一种创建和组合多个模型以获得特征表示改进的技术。现有的理论研究和经验研究表明，整体通常比整体内的任何个体分类具有更准确的分类效果。当神经网络集成用于回归估计时，集成的输出通常由各网络的输出通过简单地组合（平均或加权平均）产生。同时，Sollich 等人[126] 还指出当不同神经网络特征表示之间的差异性越大的时候，其组合而成的最终表示更倾向于产生更好的结果，原因在于其能够一定程度上去除假设偏差 (Hypothesis bias)。例如，当模型之间存在显着差异时，神经网络集成往往会产生更好的结果。因此，一个好的神经网络集成应该试图促进它所结合的模型之间的多样性。

而在行人再识别任务上，如何很好的使用神经网络集成来提升行人再识别性能也是一个值得研究的问题。最近，Zheng 等人[128] 提出了一个行人对齐网络，该模型利用空间变换网络 (Spatial Transform Network, STN)[129] 在第二个分支上应用空间转换用于处理行人图像未对齐问题，然后将来自两个分支的特征表示组合为最终行人描述符。实质上，他们的方法可以被视为网络集成的一个实例。本章所提出的方法和[128] 具有相同的精神，但在考虑引导两个分支以及自学习掩码以用于网络内的潜在特征选择方面存在差异。本章所提出的网络结构正是基于神经网络集成对于最终特征表示泛化能力具有提升的效果的事实上。更进一步来讲，良好的特征差异性是获取鲁棒的特征表示的重要基础，本章为了扩大集成网络之间的差异性也设计使用了成对排序损失。

2.4 基于互补特征提取的行人再识别

一个具有越多鉴别信息，且能够覆盖行人的不同部位的特征才是越理想的特征。目前很多工作都是致力于学习单个更加鲁棒并具有鉴别行的特征表示[85,128,130,131]。行人再识别工作的难点在于当一个人的姿势以及视角都具有很强的不确定性时，很难获得一个鲁棒的表示。针对这个问题，最近很多工作都利用了注意力模型来增强效果。注意力模型通过在不同的行人图像部分给予不同的注意力来提升行人描述的鉴别性。最常见的就是在行人图像中选取不同的注意区域并且给予其不同的权重。例如，Wei 等人[132] 首

先利用人体部位关键点来提取粗糙的行人部分，并辅之以行人全图来进行特征学习。Li 等人[133] 采取了类似的方法来首先定位一些局部具有鉴别特征的区域，然后进行行人表示的学习。Zheng 等人[128] 通过 STN[129] 来自动地学习以及变换局部区域。然而本章的工作是通过利用一个高层次的特征表示生成一个低维特征的掩码 (mask) 来给这些低维的特征重新赋予权重，从而进行注意力的转移，这样的做法的好处在于其不需要像上述非常严格的矩形区域选择，从而达到一种具有非常高灵活性的注意力区域选取。正如上一节中所描述的一样，现有的特征提取方法往往聚焦于整个行人最具鉴别性的区域。图2.2中展示了行人再识别任务中的全局特征以及局部特征的重要性。图2.2的第一行展示的是正常的输入行人图片，第二行展示的是一般单支分类网络的模型注意力区域的可视化（红色框部分表示注意力高的部分），第三行展示的是，除了第二行中比较全局的注意力之外，一些局部位置的注意力（黑色框部分）能够大大帮助提高再识别的准确性。

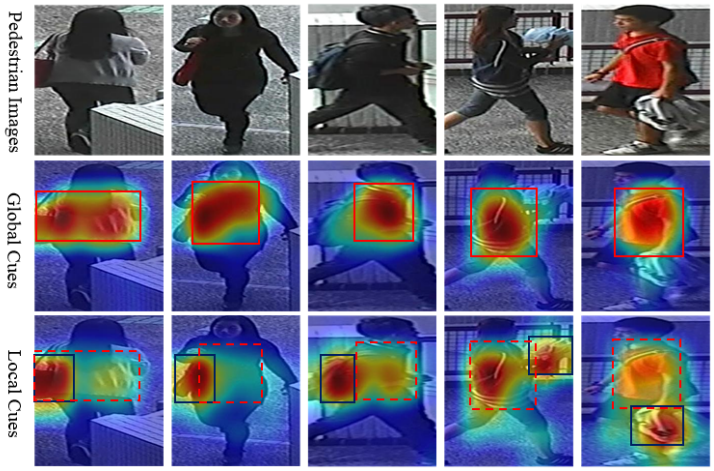


图 2.2 行人的全局与局部特征的视觉注意力区域可视化

2.4.1 特征选择网络

本章提出的特征掩码网络 (FMN)，其网络结构如图2.3所示。该网络结构由四个部分组成，分别是：（1）基础网络 (Base Network，BN），（2）全局特征网络 (Global Representation Network, GRN）（3）混合网络 (Mixing Network, MN)，（4）配对特征网络 (Counterpart Representation Network, CRN)。

基础网络 基础网络是由一个卷积层和一个最大值池化层组成，其主要功能是学习一些边角信息以及低层次的图像表示，并且为后续的网络提供数据输入。

全局特征网络 全局特征网络针对一张输入图片学习一个全局的特征表示，其设计和 ResNet-50[124] 的结构一致，由五个残渣块（Residual blocks）组成，ResNet-50 的第一个模块充当上述的基础网络。该网络的参数是预先在图像分类数据集 ImageNet 上预训练之后，再将其放到行人再识别的数据集上进行微调（fine-tuning)。

混合网络 混合网络的功能在于利用全局特征网络为低层的特征输入生成一个掩

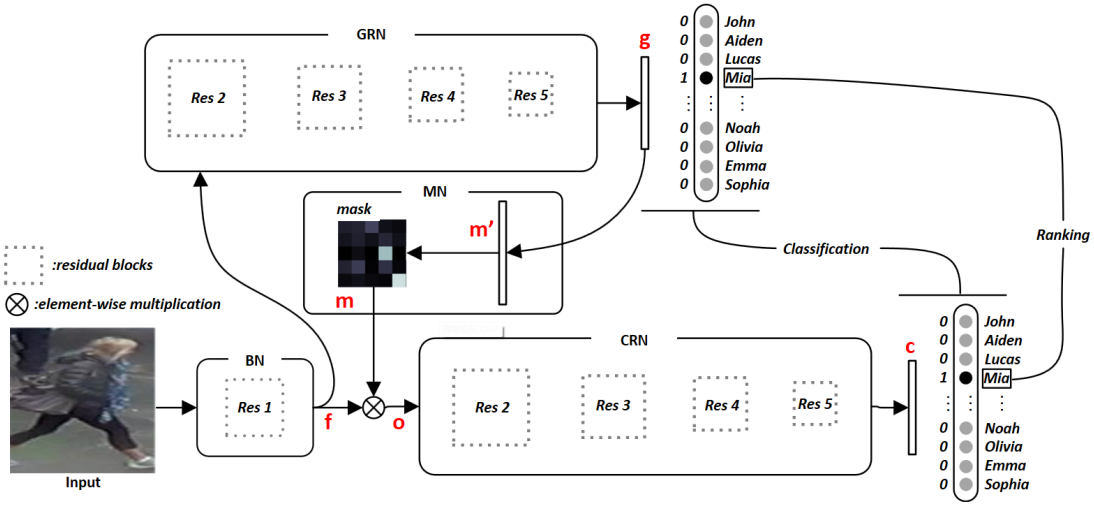


图 2.3 特征掩码网络的网络结构示意图

码。这些掩码权重是由全局特征网络中的特征表示经过一个全连接层来实现的。混合网络接受来自于全局特征网络的行人特征并使之通过一个转换器，该转换器由一个全链接层扮演，经过转换器后的特征表示又被改变至目标特征图的大小尺寸并经由一个混合器将其与低维的特征进行混合，得到的经过掩码的特征再输入到配对特征网络中。网络中的掩码可以对低维特征重新赋予权重，因此其也可以被看为是一个注意力模型，其中对配对网络来说有用的鉴别区域被给予了更多的权重。

配对特征网络 配对特征网络的功能在于学习一个与全局特征网络具有互补性的特征来完善最终的特征表示。配对特征网络的输入来自于基础网络和混合网络之后的输出。配对特征网络具有与全局特征网络相同的网络结构，但两者的参数并不共享。参数的不共享导致了其可以学习到一个不同的、具有不同侧重点的局部表示。这个表示可以作为全局特征网络表示的一个互补的信息。这在本章后面的实验中也表明了其的有效性。下面将具体介绍掩码的计算过程。

2.4.2 掩码计算过程

混合网络的输入来自两个部分，分别是全局特征网络的特征输出 $\mathbf{g} \in \mathbb{R}^m$ 和基础网络的低层特征输出 $\mathbf{f} \in \mathbb{R}^n$。由于在卷积网络中的图片的特征表示都是按照二维激活图的形式存储，所以它们的表示可以很方便地写成：$n = h \times w \times c$, 其中 h，w，c 分别表示特征高、宽和通道数。混合网络首先将全局特征网络的表示转换成一个 $n' = h \times w$ 维度的输出，该输出将会被用来做特征掩码：

$$\mathbf{m}' = \mathbf{W}^{\mathrm{T}}\mathbf{g} \tag{2.2}$$

其中 $\mathbf{m}' \in \mathbb{R}^{n'}$ 是中间掩码值，$\mathbf{W} \in \mathbb{R}^{m \times n'}$ 是转换器的权重，其也就相当于本网络结构中的全链接层。经过这样的掩码转换之后，可以直接从全局特征中学习出一个特定图像相关（image specific）的掩码来支持配对特征网络的特征学习。由于这里的目的是为了在空间维度进行局部信息的加强和抑制，所以该模型对 f 中所有的通道都进行相同的

掩码操作。最终的特征掩码 ($\mathbf{m} \in \mathbb{R}^{h \times w}$) 可以由 $\mathbf{m}'$ 根据下面的操作进行形状调整并对每个元素进行指数操作。

$$\begin{aligned} \mathbf{m}_{i,j} &= exp(\mathbf{m}_k), \\ s.t., \quad j &= \lfloor k/h \rfloor + 1, i = k - h(j-1) \end{aligned} \tag{2.3}$$

在获得到特征掩码 $\mathbf{m}$ 之后，混合网络使用一个混合器来将其与 $\mathbf{f}$ 进行融合：

$$\mathbf{o}_i = \mathbf{f}_i \circ \mathbf{m}, \quad s.t., \quad i \in [1, c] \tag{2.4}$$

其中，i 表示 $\mathbf{f}$ 中的通道数。

特征掩码的作用 本章认为通过混合网络产生的特征掩码在该网络结构中起到了两个主要的作用。

(1) 由于混合网络的输入是全局特征，所以经过转换后的掩码中自然包含了全局特征信息，这样一来，全局特征信息也能够被输入到配对特征网络中，将两个分支打通来促使配对网络进行互补信息的学习。

(2) 特征掩码是按照空间维度对特征重新赋予权重（每层通道的掩码值一样），因此其掩码本身可以看成一个隐藏的注意力模型，促使配对网络能够将注意力转移到一些需要被注意的局部区域。这一点对于互补特征的学习很重要，因为通常来讲，行人具有鉴别性的特征信息会分散在行人的不同部位，例如，鞋子和配件，而不是只是单单聚集在行人的躯干部分。

2.4.3 训练过程

本章所提出来的网络采用了分阶段、多任务训练的方法。其训练过程可参照算法1。首先，BN 和 GRN 使用在 ImageNet 数据集上预训练过的 ResNet-50 网络在行人数据集上按照行人个数进行分类训练。当这一部分的训练收敛之后，就可以通过 GRN 得到具有高层全局鉴别语义信息的特征表示。接着，固定住 BN 和 GRN 的参数不变，再次用同样的方法来对 MN 和 CRN 部分的参数进行学习。和上一阶段不同的是，通过第二步的训练，可以使得 CRN 能够关注到更多与全局高层语义信息相互补充的一些局部信息，MN 可以被看成一种软注意力模型，对 BN 和 GRN 过来的特征进行了选择，从而使得 CRN 能够与 GRN 具有差异性。

GRN 和 CRN 两个阶段的训练方式都遵循一般的分类网络的训练法方法，其可以表示为：

$$\mathcal{L}_{cls}(\mathbf{p}, \mathbf{y}) = -\sum_k y_k \log \frac{\exp(p_k)}{\sum_j \exp(p_j)}, \qquad k \in [1, r] \tag{2.5}$$

其中 $\mathbf{p} \in \mathbb{R}^r$ 表示预测激活输出 (activation output), $\mathbf{y} \in \mathbb{R}^r$ 是一个独热 (One-hot) 编码，表示预期的输出 (desired output)。在这里，r 表示的是数据集的总类别数，也相等于激活层的输出个数。

算法 1 互补鉴别特征提取网络训练过程

输入： ResNet-50 在 ImageNet 上预训练模型 $\phi(\cdot;\boldsymbol{\theta_o})$，行人再识别训练数据集 I, 行人 ID 标签 Y, GRN 和 CRN 分别的最大训练次数 T_g 和 T_c。

输出： 训练完毕的模型 $\phi(\cdot;\hat{\boldsymbol{\theta}}_b,\hat{\boldsymbol{\theta}}_g,\hat{\boldsymbol{\theta}}_m,\hat{\boldsymbol{\theta}}_c)$, $\hat{\boldsymbol{\theta}}_b$, $\hat{\boldsymbol{\theta}}_g$, $\hat{\boldsymbol{\theta}}_m$ and $\hat{\boldsymbol{\theta}}_c$ 分别表示 BN，GRN，MN 和 CRN 的学习参数。

初始化: 利用 $\boldsymbol{\theta_o}$ 初始化 $\boldsymbol{\theta_b}$，$\boldsymbol{\theta_g}$ 和 $\boldsymbol{\theta_c}$, 对 $\boldsymbol{\theta_m}$ 进行随机初始化。

全局特征学习:

for $\boldsymbol{t=1:T_g}$ **do**

固定住 MN 参数 $\boldsymbol{\theta_m}$ 和 CRN 参数 $\boldsymbol{\theta_c}$,

利用公式(2.5)训练更新学习 BN 参数 $\boldsymbol{\theta_b^t}$ 和 GRN 参数 $\boldsymbol{\theta_g^t}$,

end for

对 BN 的参数 $\hat{\boldsymbol{\theta}}_b \leftarrow \boldsymbol{\theta_b^{T_g}}$ 和 GRN 的参数 $\hat{\boldsymbol{\theta}}_g \leftarrow \boldsymbol{\theta_g^{T_g}}$ 进行更新,

配对特征学习:

for $\boldsymbol{t=1:T_c}$ **do**

固定住 BN 参数 $\hat{\boldsymbol{\theta}}_b$ 和 GRN 参数 $\hat{\boldsymbol{\theta}}_g$ 并利用 GRN 对输入进行前向运算,

利用公式(2.2)和公式(2.3)来进行掩码 $\boldsymbol{m^t}$ 的生成,

利用公式(2.4)和对 CRN 的输入进行掩码操作,

利用公式(2.5)和公式(2.6)对 MN 参数 $\boldsymbol{\theta_m^t}$ 和 CRN 参数 $\boldsymbol{\theta_c^t}$ 进行更新学习,

end for

对 MN 的参数 $\hat{\boldsymbol{\theta}}_m \leftarrow \boldsymbol{\theta_m^{T_c}}$, 和 CRN 的参数 $\hat{\boldsymbol{\theta}}_c \leftarrow \boldsymbol{\theta_c^{T_c}}$ 进行更新,

返回： 学习完毕的模型 $\phi(\cdot;\hat{\boldsymbol{\theta}}_b,\hat{\boldsymbol{\theta}}_g,\hat{\boldsymbol{\theta}}_m,\hat{\boldsymbol{\theta}}_c)$

正如前面神经网络集成部分所描述的，特征组合能有更好的表现的原因在于其之间的分支差异性，为了能够更好的扩大 GRN 与 CRN 之间的差异性，本章还提出了在两个网络之间加上一个跨支 (inter-branch) 的成对排序损失函数（pairwise ranking loss），其定义如下：

$$\mathcal{L}_{rank}(p_t^G, p_t^C) = \max\{0, p_t^G - p_t^C + m\}, \tag{2.6}$$

其中，m 代表的是边际阈值（margin），实验对此参数并不敏感，实验中设置为 1，p_t^G 和 p_t^C 分别是 GRN 和 CRN 对于目标类别标签 t 的预测激活输出值。该损失函数的加入使得 CRN 能够利用 GRN 的特征作为参考，并促使 CRN 对于正确类别标签的预测值要比 GRN 大至少一个边际阈值，从而能够得到更加准确与可信的预测。于是，本章所提出的第二支分支网络最终的损失函数可以由公式(2.5)和公式(2.6)组合而成，写成如下形式：

$$\mathcal{L} = \mathcal{L}_{cls} + \lambda\mathcal{L}_{rank} \tag{2.7}$$

其中 λ 表示的是分类损失和排序损失之间的调和参数。

2.4.4 讨论

排序损失函数选取 基于排序的损失函数一般来说有两种主要的设计方法。

(1) 第一种排序损失是在现有行人再识别领域广泛应用的方式，其被应用在深度学习网络的倒数第二层的特征表示空间中。该方式对于图像的输入有一定的要求，需要输入是二元组 (pairs) 或者三元组 (triplet) 的形式。在输入的是成对或者三元组的情况下，排序损失致力于促使同类样本之间距离越小越好，不同类别之间的距离越大越好，即起到拉近类内距离，推远类间距离的作用，从而使得提取出来的特征具有更好的鉴别性于鲁棒性。但是其有一个比较大的缺点就在于如何使用该种排序损失需要对输入数据多元组进行组织，而这样的输入组合的数量会随着数据量的多少呈指数型增长。

(2) 第二种排序损失在文献[134]中的基于孪生网络 (Siamese Network) 细粒度（fine-grained）图像分类任务中被提出，在本章中这样的损失被称为跨支 (inter-branch) 成对排序损失，其应用在深度学习网络的最后的预测激活输出上。具体来讲，使用一支的预测作为参考，要求另一支能够做出比其更加精准的预测来保证鉴别特征的挖掘，除此之外，跨支排序损失对输入没有特别的要求，因为其是应用在同一张图片的不同分支上，所以避免了第一种排序损失所面临的输入数据量呈指数增长的困境。

多损失训练 本章所提出的第二支互补特征的最终损失函数如公式(2.7)所示。其两个部分分别对应的是分类损失函数 (公式(2.5)) 和排序损失函数 (公式(2.6))，这两个部分分别对第二支分支网络的特征学习起到了不同的作用。分类损失函数能够保证所学特征有一定的鉴别性，而排序损失函数是用单支的预测值作为基准，促使了该分支对于目标类别要有更高的预测值，从而保证了该分支能够挖掘到更多的互补特征。多任务的训练方式很好的符合了本章所提出来的网络结构，并且能够充分利用分支网络之间的关系。值得一提的是，附加的排序损失函数并不代表该匹配网络分支的特征具有更高的重要性，仅仅是为了提高特征之间的差异性。而两支网络特征的重要性在最终行人描述子组合时（见公式(2.8)）则是可以通过一个权重系数来调节的。

2.4.5 行人描述子

当网络训练完成之后，最终的行人描述子就可以由 GRN 和 CRN 两路分别产生的行人特征表示组合来得到。最终的特征表示是取自于各个网络的倒数第二层的激活值，这些特征表示分别具有不同的行人鉴别信息。以下即为两支特征进行融合而获得最终行人表示的方法：

$$D = [\alpha \frac{\mathbf{g}^{\mathrm{T}}}{\parallel \mathbf{g} \parallel_2}, (1-\alpha) \frac{\mathbf{c}^{\mathrm{T}}}{\parallel \mathbf{c} \parallel_2}]^{\mathrm{T}}, \tag{2.8}$$

其中，$\|\cdot\|_2$ 操作符表示的是 ℓ^2-标准化。参数 α 决定了分别从 GRN 和 CRN 得到的特征 $\mathbf{g}, \mathbf{c}$ 之间的比重。在实验中，α 的值被设置成 0.5。最终所得到行人描述子被用来在

行人库中按照欧氏距离的大小进行最近邻搜索从而进行行人的再识别。

2.5 实验结果与分析

本章的实验所用操作系统版本为 Ubuntu14.04 LTS，使用基于 Matlab 的深度学习框架 MatConvnet[135] 来实现所提出的特征提取方法 FMN，实验显卡为 Nvidia GTX 970，显存为 8G。

2.5.1 数据集和评价标准

本章实验使用三个目前公开的大规模行人再识别数据集，分别是 Market-1501[8]，DukeMTMC-reID[95] 和 CUHK03[13]。其中值得注意的是 CUHK03 数据集在文献[95]中提出了新的训练集和测试集划分方式，以满足在实际情况下可获取的训练数据量会小于测试集的数据的设定。为了保证所提出方法的有效性和通用性，本章中所有的数据库设定与文献[128]保持一致，并用相同通用的方法来进行性能比较。

在本章中，行人再识别的评价指标采用的是平均精度均值 (mean Average Precision, mAP) 和首选识别正确率 (rank-1 accuracy)。

2.5.2 实现细节

由于 ResNet-50 在分类问题上可以取得相当高的准确率，并且参照 Zheng 等人[128] 提出的基准方法，本章也采用同样的网络架构。该网络的初始参数为在 ImageNet 上预训练过的参数。为了将其迁移到行人再识别领域，该网络最后具有 1000 个激活单元的网络层被替换成了目标数据集的行人个数个单元的激活层。值得注意的是，BN 与 GRN 或者 CRN 分支都可以分别组成一个完整的 ResNet-50 网络，也就是说，本章所提出的网络结构可以看成两个 ResNet-50 的结构，但是它们之间共享 BN 部分。在进行全局特征学习的阶段，BN 与 GRN 部分参数的学习率初始化为 0.1，20 个训练阶段 (epoch) 之后降低到 0.01。MN 和 CRN 部分参数的学习率同样也是预设为 0.1，但是在 35 个训练阶段之后衰减到 0.01。对于两个分支，在实验中都设置了可以使其收敛的最大训练阶段为 50 个训练阶段。模型中所有的参数使用的是随机梯度下降更新方法，动量设为 0.9。在实验中，通过对训练集的图像进行了随机左右镜像和随机裁剪的操作来进行数据预处理和增广。

2.5.3 对比消融实验

下面对所将提出的网络结构的各个组件（components）进行了对比消融实验（ablation study）以直观地体现出本章所提出的网络组件的贡献。

基准 (baseline) 网络 本章的基准网络采用的是 ResNet-50，并根据行人再识别的数据集进行倒数第二层网络激活单元个数的替换并进行参数微调 (fine-tuning)。基准网络在本章使用的 4 个数据集 (Market-1501，DukeMTMC-reID, CUHK03 detected 和 CUHK03 labeled) 上的首选准确率 (rank-1) 分别为 79.33%, 67.91%, 38.00% 和 34.36%。

这个基准网络已经取得了比较好的准确率，算是一个比较有挑战性的性能基准。

GRN 对比 CRN GRN 分支和 CRN 分支的单独的性能结果可见表2.2。可以看出的是，单独的 CRN 的性能结果都很接近或者稍小于 GRN，推测的原因是因为其更多的挖掘的是互补的特征，所以单独使用其作为最终特征表示会导致一些性能下降，这种情况可以在图2.2中得到更好的解释，图中的红色实线框分别对应着不同图片 GRN 分支的注意力所在位置，可以看出来的是在 CRN 的注意力图中，相同位置（由红色虚线框标出）的 CRN 特征的注意力程度有所下降，而其余的由黑色方框框出来的部分的注意力程度显著增强。这也符合并证实了本章所提出的网络结构的两个不同分支能够对图像的不同特征进行提取，并且两者有着相当程度的互补性，从而也揭示了在节2.4.5中选取将两支的特征进行拼接作为最终行人描述子的原因。值得注意的是，CRN 在 CUHK03 detected 数据集上的性能比 GRN 低了 6.5%，其潜在的原因是模型的过拟合（overfitting）问题，因为实验中对于所有的数据库采用了同一的模型参数，而 CUHK03 数据集只有 DukeMTMC-reID 数据集大概一半的训练数据量。即便如此，FMN 的性能提升还是可以表明 CRN 学到了于 GRN 相互补的特征。

表 2.2 特征掩码网络的消融比较实验

方法	维度	Market-1501				DukeMTMC-reID			
		rank-1	rank-5	rank-20	mAP	rank-1	rank-5	rank-20	mAP
GRN	2048	79.33	91.48	96.62	58.50	67.91	81.33	89.59	48.40
CRN$^-$	2048	80.28	90.07	94.71	59.23	68.88	82.05	89.86	50.62
CRN	2048	81.15	91.39	96.44	59.88	69.84	83.03	90.04	51.05
FMN$^-$	4096	85.12	93.08	96.57	65.93	73.61	**85.19**	92.06	56.55
FMN	4096	**86.00**	**93.74**	**97.51**	**67.12**	**74.51**	85.05	**92.41**	**56.88**

方法	维度	CUHK03 detected				CUHK03 labeled			
		rank-1	rank-5	rank-20	mAP	rank-1	rank-5	rank-20	mAP
GRN	2048	38.00	48.14	59.93	33.68	34.36	46.64	58.14	30.14
CRN$^-$	2048	32.71	45.03	59.00	29.69	31.29	47.04	61.27	30.22
CRN	2048	31.50	46.07	60.86	29.39	32.36	47.57	61.93	30.41
FMN$^-$	4096	40.79	53.00	64.57	37.94	39.55	53.79	65.20	37.66
FMN	4096	**42.57**	**56.21**	**67.36**	**39.21**	**40.71**	**54.57**	**65.50**	**38.05**

排序损失 跨支的排序损失函数在考虑到两支之间的关系的基础上更加进一步的扩大了两支特征之间的差异性并有利于最终的性能提升。表2.2中带有 $^-$ 的表示没有加入排序损失函数的模型，即其最终训练损失函数只有分类损失函数，而其余的均表示是

采用的2.4.4中所描述的多损失函数训练。从表中可以看出，当排序损失函数加入网络之后，行人再识别的性能普遍有所提升，这表明排序损失函数的加入有助于行人再识别任务的进行。但是，其所取得的性能提升相较于互补特征所导致的性能提升来说是相对比较小的，平均的性能提升在 1%-2% 左右。

表 2.3 Market-1501 数据集不同 BN 结构实验结果

层	掩码尺寸	rank-1	mAP
pool1	56×56	**86.00**	**67.12**
conv2_x	56×56	85.15	65.84
conv3_x	28×28	84.23	65.16
conv4_x	14×14	82.39	61.95

低层特征选择 卷积神经网络能够针对不同特征尺度进行信息提取，从初始层到最终层，所提取的特征分别对应着低层特征和高层特征。本段将对 BN 的不同特征尺度对行人再识别的性能影响进行评估。表2.3中列举了不同的 BN 网络结构选择在 Market-1501 数据集上对行人再识别性能的影响。正如在章节2.5.2中所述，ResNet-50 被分成两个部分，分别是 BN 与 GRN，为了能够简单地对研究网络的分割进行表示，实验中对网络结构中的层采用 ResNet-50 原本的命名方法（例如，pool1，conv1_x，pool5）。从表2.3中可以明显看出，当 BN 所选的层越靠近初始层，行人再识别的准确率会越高。具体来讲，conv4_x 层的 rank-1 准确率和 mAP 分别是 82.39% 和 61.95%，而 pool1 层的 rank-1 准确率和 mAP 分别是 86.39% 和 67.12%，有了 4-6% 的性能提升。这个现象也是符合经验规律的，因为最终的行人描述子是通过拼接而成，而两支的输入能够分别包含更多的细节特征的基础上，所学的特征才会有更多的差异性。当 BN 的层数越来越深时，GRN 和 CRN 的输入特征已经比较高层次偏语义，会导致细节不够，从而导致特征差异性无法保证。越是低层的输入特征越能够帮助网络判别一些细微的外观差距。当 BN 选择的特征越低层，其局部信息越多，从而能提升最终再识别性能。

相近工作比较 与本章工作最为相近的一篇近期的工作是文献[128]，但是本章与该工作的特征学习方法以及目标都有比较大的差别。文献[128]着力于解决自动检测的行人图片存在的对齐问题，其通过使用一个网络自动学习仿射变换 (affine transformation) 参数来对图像或特征图进行变换。与之不同的是，本章的工作主要聚焦在互补特征的学习与提取，其原因在于一些相对于比较局部特征对于整体特征来说具有一定的互补性，将互补的特征作为最终行人描述子能够有助于行人再识别任务的进行。本章所提出的方法不需要对图像进行变换，通过一种更加简单的注意力模型就可以得到很好的效果。除此之外，本章提出了一个多损失训练能够尽可能扩大两支表示之间的差异性，并且分别学到全局和局部的行人鉴别特征。总而言之，本章的工作相于比 PAN[128] 差别在于：1) 文献[128]可以理解为文献[129]的一种变型，使用在了行人再识别领域来解决行人图片的对齐问题；2）本章所提出的网络除了网络结构与 PAN[128] 不同，还提出了一个多损

表 2.4 特征掩码网络与神经网络集成的比较实验

方法	维度	Market-1501				DukeMTMC-reID			
		rank-1	rank-5	rank-20	mAP	rank-1	rank-5	rank-20	mAP
GRN	2048	79.33	91.48	96.62	58.50	67.91	81.33	89.59	48.40
GRN*	2048	79.42	90.23	95.88	58.35	68.02	81.46	89.66	49.02
CRN	2048	81.15	91.39	96.44	59.88	69.84	83.03	90.04	51.05
NE	4096	82.00	92.27	96.34	62.68	71.19	83.98	91.94	53.23
FMN	4096	**86.00**	**93.74**	**97.51**	**67.12**	**74.51**	**85.05**	**92.41**	**56.88**

方法	维度	CUHK03 detected				CUHK03 labeled			
		rank-1	rank-5	rank-20	mAP	rank-1	rank-5	rank-20	mAP
GRN	2048	38.00	48.14	59.93	33.68	34.36	46.64	58.14	30.14
GRN*	2048	37.17	47.28	58.82	33.71	33.98	46.25	57.89	29.86
CRN	2048	31.50	46.07	60.86	29.39	32.36	47.57	61.93	30.41
NE	4096	40.33	50.45	62.56	35.74	36.68	49.02	62.58	34.14
FMN	4096	**42.57**	**56.21**	**67.36**	**39.21**	**40.71**	**54.57**	**65.50**	**38.05**

失的训练方法，使得该模型的两支特征能够有更大的差异性；3）本章的方法在所有的数据集上在同样的实验设定下都取得了比文献[128]要高 5-6% 的 rank-1 准确率和 mAP 值。

2.5.4 结果评价

特征掩码网络的有效性 本节将在三个广泛使用的行人再识别数据集上进行实验来证明本章所提出的特征掩码网络对于行人再识别任务的有效性。实验结果均在表2.2中展示出来。可以看出的是，当 GRN 和 CRN 两支的特征组合成最终行人特征表示的时候（FMN 行），行人再识别的准确率有了很明显的提升。具体来讲，Market-1501,DukeMTMC-reID 和 CUHK03 数据集上的 rank-1 准确率相对于基准网络分别提升了 6.66%, 6.60%, 4.57% 和 6.35%。mAP 也得到了一个很明显的提升，分别为 8.62%, 8.48%, 5.53% 和 7.91%。在多个数据库上的一致性能提升表现充分全面证明了本章所提出的特征掩码网络对于行人再识别任务的有效性。

NE 对比 FMN 如节2.3中所述，本章所提出的特征掩码网络可以看成是一种应用了网络集成技术的实例。表2.4中的 GRN 和 GRN* 即为两支分别单独训练的网络，满足神经网络集成的训练条件。可以看到的是，GRN 和 GRN* 在所有数据集上的表现都很接近，这也是符合预期的，由于两支网络的结构完全一模一样，仅仅存在一些参数初始化的差异而已。表中的 NE 行表示的是将 GRN 和 GRN* 的特征进行拼接之后进行

表 2.5 Market-1501 数据集实验结果

方法	rank-1	mAP	方法	rank-1	mAP
DADM[83]	39.41	19.62	DeepTransfer[136]	83.69	65.53
BoW+kissme[8]	44.42	20.76	GAN[116]	83.97	66.07
MR-CNN[137]	45.58	26.11	PAN[128]	82.81	63.35
MST-CNN[138]	45.10	-	APR[139]	84.29	64.67
FisherNet[60]	48.15	29.94	Triplet[140]	84.92	69.14
CAN[81]	48.24	24.43	PAN+re-rank[128]	85.78	**76.56**
SL[40]	51.90	26.35	JLML[141]	85.10	65.50
DNS[19]	55.43	29.87	DPFL[142]	88.90	73.10
Gate Reid[80]	65.88	39.55	GLAD[132]	**89.90**	73.90
SOMAnet[143]	73.87	47.89	HA-CNN[144]	**91.20**	75.70
PIE[145]	78.65	53.87	Baseline	79.33	58.50
Verif.-Identif.[146]	79.51	59.87	FMN	86.00	67.12
SVDNet[147]	82.30	62.10	FMN+re-rank	87.92	**80.62**

行人再识别的性能结果。结果表明，尽管两支网络 GRN 和 GRN* 是分开训练，而且单独的性能差不多，当两者结合在一起的时候可以发现，最终的性能均有了 2-3% 的提升。然而，当使用本章提出的特征掩码网络将两支进行关联之后，所得到的性能 (FMN 行)，可以观察到，所有的 rank-1 指标在三个数据集上又再往上分别提升了 4.00%, 3.32%, 2.27% 和 4.03%, 同时 mAP 也有了 4% 左右的提升。这些一致的性能提升表现也充分说明了本章所提出的网络结构的合理性和有效性。

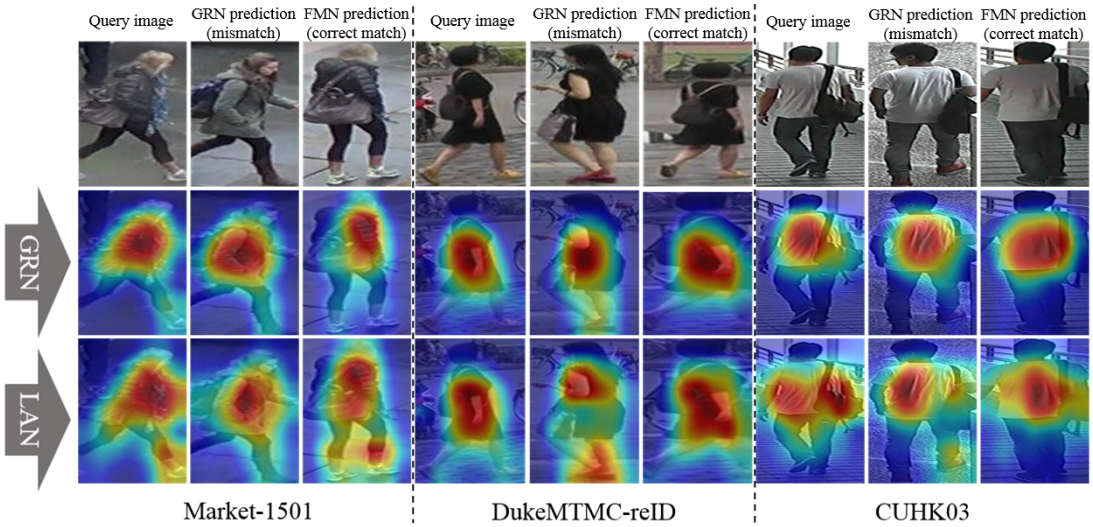


图 2.4 特征掩码网络的视觉注意力区域可视化

结果可视化 除了以上的在行人再识别数据集上大量的性能指标上的分析比较之

表 2.6 DukeMTMC-reID 和 CUHK03 数据集实验结果

方法	DukeMTMC-reID		CUHK03 Detected		CUHK03 Labeled	
	rank-1	mAP	rank-1	mAP	rank-1	mAP
BoW+kissme[8]	25.13	12.17	-	-	-	-
BoW+XQDA[8]	-	-	6.41	6.41	8.02	7.31
LOMO+XQDA[12]	30.75	17.04	12.8	11.51	14.82	13.61
ResNet+XQDA[95]	-	-	31.10	28.22	32.03	29.62
PAN[95]+re-rank	-	-	34.72	37.42	38.11	40.30
GAN[116]	67.68	47.13	-	-	-	-
OIM[148]	68.12	-	-	-	-	-
APR[139]	70.71	51.90	-	-	-	-
PAN[128]	71.61	51.51	36.32	34.01	36.90	35.00
PAN+re-rank[128]	75.92	**66.71**	41.90	**43.82**	43.91	**45.81**
DPFL[142]	79.22	60.61	40.71	37.02	43.51	40.50
HA-CNN[144]	**80.5**	63.80	41.72	38.64	**44.44**	41.00
Baseline	67.91	48.40	38.21	34.05	34.41	30.13
FMN	74.51	56.88	**42.6**	39.22	41.02	38.10
FMN+re-rank	**79.52**	**72.79**	**47.52**	**48.50**	**46.03**	**47.61**

外，本节也将采用一种能够更加直观的方式对两支网络特征的互补性做出一些展示。图2.2和图2.4中分别展示了分别从三个数据库中挑选的查询图片，并将其通过基准网络和本章提出的特征掩码网络的首选匹配图片列出来，所选图片的基准网络首选匹配均为错误匹配，而通过特征掩码网络的首先匹配为正确匹配。与此同时，图2.4也对 GRN 和 CRN 分支网络所获得的特征注意力区域进行了可视化，颜色越鲜艳代表注意力程度越高。可以看出的是，GRN 对于不同的图像的注意力都主要集中聚集于行人身躯中段的部位，也就是说，GRN 网络更加倾向于学习一个全局的高层语义的表示。而 CRN 则是在保留一些全局语义表示的情况下去挖掘出了更多的局部特征。具体来讲，对于第一个例子来说，GRN 提取的三张特征都聚集在行人的躯干部位，查询图与错误匹配在注意力位置和大小上都有很大的相似，而正确匹配的注意力范围的位置和大小则有比较大的差别，这也就解释了为什么该错误匹配会出现的原因。然而，CRN 提取的三张特征图却有了不同的着重点，可以看到 CRN 在查询图和正确匹配图的注意力都有一部分放在了右脚踝的部分，其右脚踝并没有被裤子覆盖，而错误匹配的行人的穿的却是长裤，这也就是两者比较明显的具有区分的特点。这些特点往往都是在一些比较局部的地方或者散落在图片的各处，而有了 CRN 这样能够挖掘到这些局部的特征，无疑能够帮助行人再识别的进行。类似地，第二张和第三张图片的差别都在于错误匹配的人穿着了红色

的鞋子而查询图的人没有。除此之外，下面还列出一些查询样本的匹配结果示意图，见

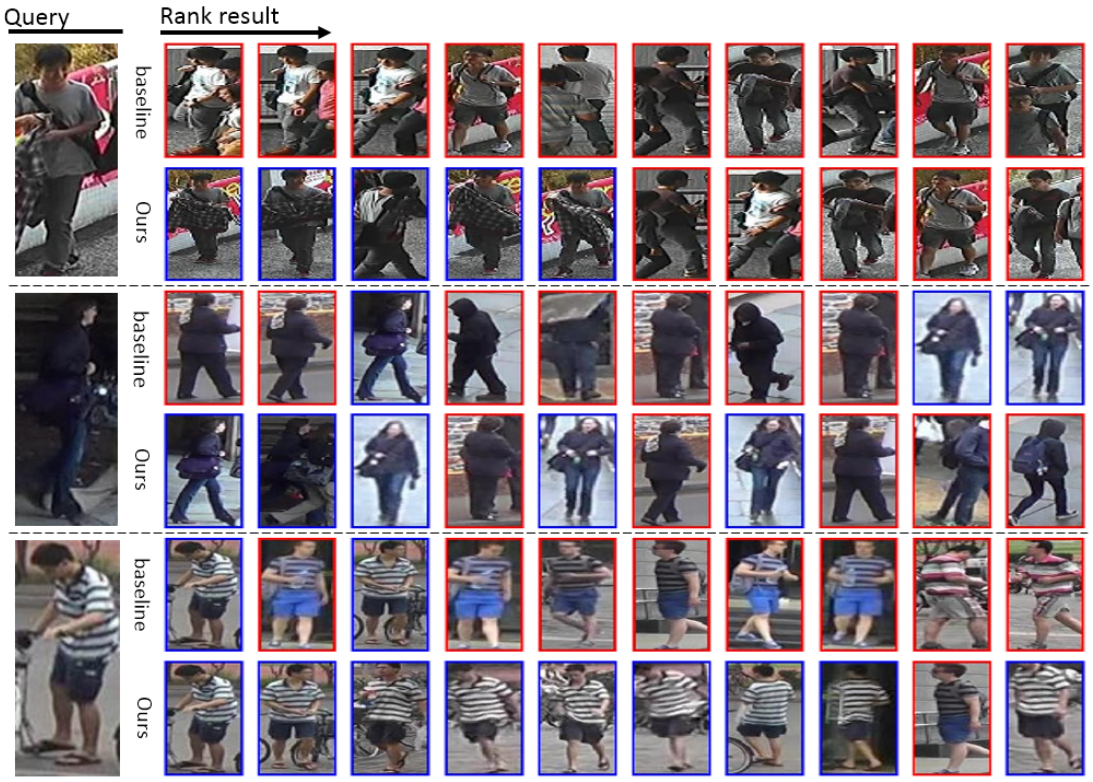


图 2.5 不同数据集上的行人再识别结果检索列表

图2.5。图中红色框表示与查询图片行人的错误匹配，而绿色框表示与查询图片行人的正确匹配，匹配图片从左到右按照相似度由小到大进行排序，对应于每个查询图片，上面一行为基准网络的行人匹配列表，下面一行为 FMN 的行人匹配列表。由图可见基准网络的识别性能比较低，其很容易与外形相似的人进行混淆，而经过了互补特征的学习之后，FMN 能够很大程度上提升其行人再识别的准确率。

与同期研究比较 下面将进行本章提出的特征掩码网络与其他方法的比较，其中包含高水平的行人再识别方法。本节分别在表2.5和表2.6中给出在 Market-1501，DukeMTMC-reID 以及 CUHK03 数据集上的性能比较。总的来说，本章所提出的模型在所有的方法中所采用的网络结构是相对来说很简单的，但是却取得了相当具有竞争力的结果。具体来说，图2.5中给出了在 Market-1501 数据集上的一些方法。本章所提出的特征掩码网络取得了 86% 的 rank-1 准确率和 67.12% 的 mAP 值。再通过文献[95]所提出的再排序 (re-ranking) 方法，特征掩码网络最终的效果达到了 87.92% 的 rank-1 准确率和 80.62%。这样的结果相较其他方法来说，在两个评价指标上都去的了比较一致的性能提升。DPFL[142] 在 rank-1 准确率上比本章提出的方法高出 1%，具体原因在于以下几点：1）文献[142]所选取的基础网络比本章所选择的 ResNet-50 有更强的表示能力，其表现在基准网络的性能指标。本文的基准网络所取得的 rank-1 准确率在 79.3%，相较于文献[142]的 83.3%，低了有 4%；2）DPFL[142] 有着更加复杂的网络结构，其使用了多尺度的网络结构在进行训练，对比于本章的单尺度网络对有尺度差异的图像有着一定的

优越性。GLAD[132] 在 rank-1 准确率上比本章的方法高出了 2%，原因在于其使用了一个四支网络来分别学习局部和全局的不同尺度的信息，并且这些分支进行了参数共享，避免了由于参数过多而可能导致的过拟合现象。在 DukeMTMC-reID 数据集上，本章所提出的特征掩码网络也取得了很好的性能，rank-1 准确率和 mAP 值分别是 79.52% 和 72.79%。HACNN[144] 取得了最高的 rank-1 准确率原因在于其引入了更多维度的注意力机制，包含特征的空间维度和通道维度，从而可以得到一个更加健壮的行人表示。在 CUHK03 detected 和 labeled 数据集上，本章所提出的特征掩码网络的 rank-1 准确率分别比排名第二的 PAN[128] 要高出 5.6% 和 1.6%。在 labeled 数据集上的性能提升稍微弱一点的一部分原因在于 labeled 数据集本身的特性，其行人标注均由人工手动标注，所以其标注质量较高，不会存在一些例如标定框不准确、标定框尺寸偏差等问题。

2.6 本章小结

本章基于全监督的行人再识别设定下，提出了一种能够自主地进行互补鉴别特征提取的网络结构，称为特征掩码网络 (FMN)，一些经验研究表明单支的神经网络往往只能学习到整体的高语义的特征，然而鲁棒的行人鉴别特征需要通过多处多层次的特征组合而成，本章节从此入手提出并验证了通过有特征掩码的双支网络能够同时学习全局和局部的特征，并且通过排序损失进一步地扩大了两者之间的差异性以获得更好的行人再识别结果。最后通过注意力可视化的方法直观地表现出本章所提出网络的效果，并证明了两支特征聚合是有意义且对识别性能提升是有效的。不同数据集上的实验结果均表明本章所提出的互补特征提取算法相较于现有的大部分方法具有一定的优越性。

3 基于伪数据伪标签正则化深度模型的行人再识别

目前对于行人再识别的任务，大多数的工作是集中在类似前章中的全监督机制下鲁棒特征的学习方法的研究之上。因此，大量的方法在学习过程中都是需要大量行人数据以及标签作为数据支持的。然而实际情况下，即使是现有的大规模的行人再识别数据集，其中包含的行人数目以及图片数量，相较于现实生活中的应用来说都是远远不够的。行人再识别数据的收集和标注是十分昂贵的，因为其任务特性并不像图像分类，其涉及到行人之间的关联问题。因此，如何在现有的数据的基础上，采用一些数据增广的方式来对数据集进行扩充，并且能够提升行人再识别的性能成为了一个值得解决的问题。在行人再识别中，存在着大量未标注的无监督数据。在监督学习范式下的方法中，这些无监督数据并没有被利用。如何从无标签的数据中挖掘到有利于模型优化的信息，从而增强 CNN 模型的泛化能力，得到鉴别能力更强的 CNN 模型，是本章所研究的关键问题。因此，本章基于 GAN 网络的图像生成能力，以及生成图像与原始图像之间的相关性，提出了一种半监督设定下的基于特征相似性的伪标签生成方法（Feature Affinity based Pseudo Labeling, FAPL），并利用其对伪数据生成标签从而实现模型正则化的方法。

3.1 引言

行人再识别领域在近年内的最新进展可归功于以下两个因素：(1) 使用卷积神经网络进行行人特征表示的学习[82,136]。(2) 大规模行人再识别数据集的可用性[4,8,13,116]，即大量的数据标注工作的进行，是有效地进行特征学习的一个必要条件。是由于其繁琐和劳动密集的工作性质，手动标注是获取大规模注释的主要瓶颈。这个问题在行人再识别中尤其突出，因为行人标注涉及手动目标选择和对来自不同相机的各种角度，光照，遮挡和身体姿势的变化下的行人图像进行身份关联；在各个数据集中，每个行人 ID 的图像数量仍然非常有限，如图3.1所示。在 Market-1501[8] 中，每个行人 ID 有 17.2 个图像，CUHK03[13] 中大约有 9.6 个图像，在 DukeMTMC-reID[116] 中平均每个行人 ID 有 23.5 个图像。从图中可以看出，这些行人图像之间缺少“多样性”。因此，在 CNN 模型的训练过程中，会遭遇较高的“过拟合”风险，继而会影响行人再识别的性能。因此为了避免模型的过拟合，使用其他数据或者一些其他的数据增强的方法就显得尤为重要。在现有的很多工作中，训练中使用的行人图像通常由训练集提供，并没有任何的扩展。因此，使用无标签图像的较大训练集是否会带来任何额外好处尚不得而知。因此，实行智能数据增加来扩展训练集是很重要的。这一观察激发了 Zheng 等人[116] 采用 GAN 样本来扩大和丰富训练集。

行人再识别是在多媒体内容分析领域的一个仍具挑战性的视觉理解任务。它在包括搜索与检索，跨相机追踪和视频总结在内的多个领域有广泛应用。给定一个查询图像，

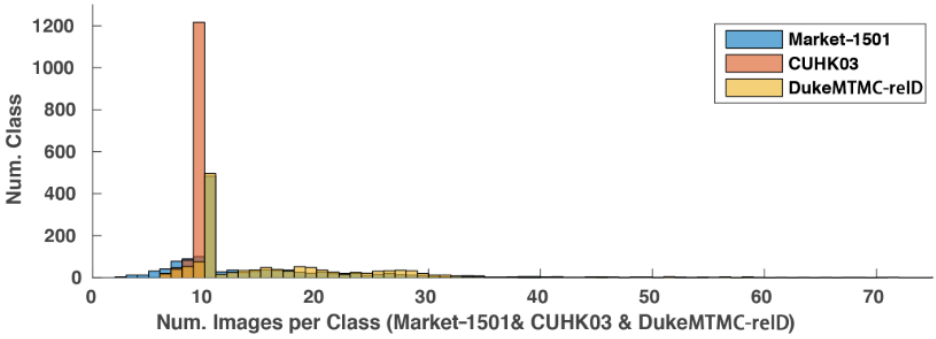


图 3.1 三大数据集的样本数量分布

它通过在一个图库中来搜索出与查询图像中为同一人的图像。这个图库通常包含来自不同相机，视角和在不同的时间点的数据。光照、物体外观、身体姿势、成像噪声、模糊和遮挡等方面的显著差异会导致较大的类内差异和较小的类间差异。本章尝试利用无监督方法生成数据作为“伪正样本”，来增强训练集中样本的“多样性”，从而减少在 CNN 模型的训练过程中存在的“过拟合”风险。其中，“伪正样本”定义为：视觉上与原始训练集中的样本有一定相似性，但属于不同行人，来源于外部独立的无监督数据集。

生成式对抗网络 (Generative Adversarial Network, GAN)[149] 的出现为这些无标签数据的获取提供了一种新的思路和方法，因为他们可以生成具有良好质感和品质的图像。然而，一个棘手的问题是如何最优地使用这些合成的数据。为它们提供“伪标签”就是一种方法，它指的是一种为无标签数据生成一些可以使其被用来训练的虚假标签的方法，它们之所以叫“伪”标签是因为其并不是一定是反应真实图像中行人 ID 的。目前在这方面已有一些工作做出了初步的尝试，也提出了一些比较简单的方法，例如 Odena 等人[150] 提出为所有生成的图像创建统一的新标签；而 Salimans 等人[151] 则建议利用从一个训练好的卷积神经网络模型中导出的类别的预测概率最大值所属的类别来为图像生成伪标签。最近，文献 [116,152] 则基于标签平滑正则化 (Label Smooth Regularization, LSR) 做一些改进为无标签的数据样本打上伪标签。LSR 是在几十年前就被提出的，最近 Szegedy 等人[153] 又对其进行了新的探究，具体来讲,LSR 的意思是在于通过分配小的值（代替 0）给非目标类进行交叉熵损失计算来减少过拟合。随后，Zheng 等人[116] 将 LSR 扩展到异常样本 (LSRO) 的情况，通过分配均匀分布的伪标签给 GAN 网络生成的图像。这个选择是为了避免直接地将所有生成的样本分类到任何一个现有的类别中去。随后，Huang 等人[152] 讨论了生成的图像之间就具有相当大的视觉差异，因此平均的标签分配并不合理，给所有图片分配相同的标签会导致模棱两可的预测这一问题。因此，他们提出伪标签的生成应该根据样本在所有预定义的类上的类别预测 (概率估计) 的排名来产生。

所有现有伪标签生成方法的一个显著缺点是它们对于无标签和有标签的数据样本之间的潜在关系是漠不关心的。这个方向上最成熟的工作[152] 采用的标签生成方法是完全依赖于输入样本（无标签的输入图像）和目标类别之间的概率预测值，而忽略了其与

真实样本之间的相似性信息。在本章的工作中，将试图通过在训练过程中将无标签的样本与真实有标签样本之间进行动态关联来克服这个缺点。受到聚类方法的启发，本章提出伪标签的生成应该利用训练数据中的潜在规则，从而提出了一种新型基于特征相似性的标签分配方法 (Feature Affinity based Pseudo Labeling, FAPL)，它能够为行人再识别模型提供显著的性能提升。FAPL 首先能够将属于同一类的数据样本在特征空间中聚集到一个比较紧密的区域中，并同时利用无标签数据与这些真实样本形成的小的簇的相似度来生成伪标签。除此之外，本章还提供除了两种可能的伪标签编码模式，即基于非最大抑制的独热伪标签和用于软标签分配的分布式伪标签。前者更容易实现和训练，而后者在本章的实验中验证其性能表现更好。

另一个对该工作有启发的观察是，尽管分别有工作从概率预测中得到的独热编码格式的伪标签和分布式编码的伪标签，但它们无法整合在一个统一的框架结构内。最开始的工作仅仅考虑了单一模式的编码，独热编码或是分布式编码。具体来说，文献[150,151,154]为无标签的数据选择了独热编码标签，而文献[116,152]则是建议采用分布式编码的标签。这种情况背后的原因是，卷积神经网络的有效训练过程是建立在有效的权重梯度回传的基础上，然而，当使用概率预测直接作为分布式编码的标签时，最终的损失函数并不能提供任何权值纠正。Huang 等人[152] 对于这一问题的解决思路是利用概率排名的来进行标签的分配，在某种程度上有效但是不可避免地引入了一些误差。在本章的工作中，通过引入特征关联来解决这个问题。特征相似性本身与类别概率没有任何关系，正是这样的无关性才使得它具有提供有效的梯度修正，从而具有统一地生成两种编码模式的伪标签的能力，并且确保参数学习过程的有效性。

考虑到最近对抗网络的进展，本章还研究了使用不同的图像生成模型对于最终结果的影响。最近有一些工作致力于提高合成图像的视觉质量，稳定模型训练过程[149,151,155–159]。因此，很多工作中对如何设计更好的损失函数以及新颖的网络框架来生成逼真的图像都做了很大的贡献。Mirza 等人[155] 提出在 GAN 的生成器和判别器都加入额外信息作为条件，如类标签，以提高生成的样本的视觉质量。在本章的工作中，除了使用 DCGAN 网络作为图像生成器，还进一步探究了利用最新改进的 Wasserstein GAN (IWGAN) 模型生成样本对结果的影响。IWGAN 模型可以避免模型坍塌 (model collapse) 并具有更好的收敛性和产生高品质样品的能力。该 WGAN[158] 是采用了更适合描绘分布之间的距离的 Wasserstein 距离度量作为对抗训练损失函数。本章的实验会显示用高质量图片能帮助提高行人再识别性能。

这项工作的主要贡献总结如下：

(1) 提出了一种半监督学习的多任务损失公式，该公式具有两个优点。首先，它联合考虑了特征空间中的类间和类内变化，以获得更具鉴别性的特征表示。其次，它可以同时为无标签数据提供伪标签。

(2) 提出了利用 GAN 生成的样本和真实数据之间的特征相似性作为伪标签的生成依据，而不是使用预测概率，并提供两种可能的编码模式来生成伪标签。此外，两

种编码模式都被证明由特征相似性统一生成。

(3) 本章所提出的方法在三个大规模行人再识别数据集上进行的实验表明，与其他伪标签方法相比，所提出的方法都取得了最好的效果。

3.2 模型正则化

卷积神经网络（CNN）的泛化能力也同样是神经网络领域研究的重点问题之一。当模型过于复杂时，例如与训练数据中包含的信息（一般来讲表现在其样本数量上）相比来说具有太多可学习的参数（表达能力过强），可能会发生过拟合 (over-fitting) 的现象并削弱其泛化能力。那么这个通过学习得到的深度模型可能描述的是随机误差或噪声而不是潜在的真正的数据分布[160]。在不好的情况下，CNN 模型可能在训练数据上表现出良好的性能，但在预测新数据时会出现很多的失误。模型正则化 (Regularization) 是防止 CNN 模型训练发生过拟合的一个关键方法。现有的很多工作已经提出了各种正则化方法, 这些方法可以大体分为传统方法[10,28,161–164] 和深度学习方法[116,165]。

3.2.1 传统数据增强

降低 CNN 模型训练过程中的“过拟合”风险有很多途径，如：在网络中减少或者共享一些参数[166]、提前停止训练过程[167] 等。近年来，基于正则化 (regularization) 的方法被广泛研究，用于降低“过拟合”风险，其中以数据增强 (Data Augmentation) 为代表。数据增强是一种显式的正则化形式，也广泛用于深度 CNN 的训练[10,124,168,169]。它旨在使用各种图像变换，例如仿射变换，旋转，翻转，裁剪，添加噪声等，从现有训练数据中人工地扩大来增加训练数据的数量，使得训练样本集更加丰富，同时样本间的多样性也得到加强。深度 CNN 训练中两种最流行和最有效的数据增强方法是随机翻转[168](flipping) 和随机裁剪 (Random Cropping)[10]。随机翻转随机地水平翻转输入图像，而随机裁剪从输入图像中提取随机子块。作为类似的选择，随机擦除[169] 可能会丢弃对象的某些部分。相较于随机裁剪可以裁剪掉对象的角落，随机擦除可能会遮挡对象的某些部分，但随机擦除维护对象的全局结构。此外，它可以被视为向图像添加噪声。随机裁剪和随机擦除的组合可以产生更多种训练数据。最近，文献[170]通过 Fast-RCNN[171] 检测来进行对抗学习，通过在空间上阻挡一些特征图来动态创建一些困难样本。随机擦除不是在特征空间中生成遮挡示例，而是通过非常少的计算从原始图像生成图像，该计算实际上是无计算的，并且不需要任何额外的参数学习。在卷积神经网络模型的学习方面也有了很多其他工作提出了许多数据增强方法，例如随机失活 (dropout)[172] 和批量归一化 (batch normalization)[173]。Dropout[10] 在训练期间以一定的概率随机丢弃隐藏神经元的输出或将其设置为零，并且仅考虑剩余权重在前向传递和反向传播中的贡献。随后文献[161]提出了一个普适性的 Dropout 方法，名叫 DropConect，它在训练期间随机选择权重并设置值为零。此外，文献[162]提出了自适应丢失 (adaptive dropout)，其中通过二元信念网络 (binary belief network) 来预测每个隐藏神经元的置零概率。Stochastic

Pooling[163] 在训练期间从多项分布中随机选择激活，这是无参数的，并且可以和其他正则化技术组合使用。PatchShuffle[164] 随机地对每个局部图像块内的像素进行重新洗牌，同时保持与原始图像块几乎相同的全局结构，这样做能够产生更多的局部变化来优化 CNN 的训练。除以上的处理方式之外，还有一些特殊的数据增强方法，如：噪声标签 (Noisy Labeling)[28,174]。最近，通过在损失层 (loss layer) 添加噪声引入了名为“DisturbLabel”[28] 的正则化方法。在每次训练迭代期间，DisturbLabel 会将小部分样本的类别标签随机更改为不正确的值。该方法通过在训练集中混入一些视为“噪声”的样本或置乱一部分训练样本标签的策略，来丰富训练集，使其泛化能力更强。以上基于正则化的方法已在深度学习模型的优化以及图像分类等领域中证明了其有效性。

3.2.2 生成对抗网络

深度学习中最引人注目的成功主要是来源于鉴别性模型，通常是那些将高维度的输入映射到类标签的模型[10]。这些具有突破性的成功主要是基于反向传播和 dropout 算法，或者使用分段线性单元[175–177]，因为这些单元具有良好的梯度特性。而深度生成模型的影响则相对来说较小，主要是因为它难以对在最大似然估计和相关策略中出现的许多难以处理的概率进行估计，并且也难以在生成环境中平衡利用分段线性单元的优点。于是 Goodfellow 等人[149] 在 2014 年提出了一种新的生成模型估计程序，来避免这些困难。在他们所提出的对抗性网络框架中，生成模型与对手进行对抗：通过一个判别模型来学习和判断输入样本是来自模型分布还是真实的数据分布。

GAN 的核心思想源于博弈论的纳什均衡。设定参与游戏的双方分别为一个生成器 (Generator) 和一个判别器 (Discriminator), 生成器捕捉真实数据样本的潜在分布, 并生成新的数据样本；判别器是一个二分类器，判别输入是真实数据还是生成的样本。为了取得游戏胜利, 这两个游戏参与者需要不断优化, 各自提高自己的生成能力和判别能力, 这个学习优化过程就是寻找二者之间的一个纳什均衡。

为了学习生成器在数据 x 上的分布 p_g，首先定义一个输入噪声的先验变量 $p_z(z)$，然后将它到一个数据空间的映射表示为 $G(z;\theta_g)$，其中 G 是由包含参数 θ_g 的多层感知器的一个可微分函数。此外，再定义了第二个多层感知器 $D(x;\theta_d)$，它的输出是一个标量。$D(x)$ 表示 x 来自数据而不是 p_g 的概率。通过对 D 进行训练以最大化为训练样本和来自 G 的样本分配正确标签的概率。同时通过对 G 进行训练以最小化 $\log(1-D(G(z)))$：

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim p_{data}(x)}[\log D(x)] + \mathbb{E}_{z\sim p_z(z)}[\log(1-D(G(z)))] \tag{3.1}$$

其中的生成器和判别器可以用任意可微分的函数，这里可采用可微分函数 D 和 G 来分别表示判别器和生成器，它们的输入分别为真实数据 x 和随机变量 z。$G(z)$ 为由 G 生成的尽量服从真实数据分布 p_{data} 的样本。G 的目标是使自己生成的伪数据 $G(z)$ 在 D 上的表现 $D(G(z))$ 和真实数据 x 在 D 上的表现 $D(x)$ 一致。这里的 D 的目标是实现对数据来源的二分类判别。如果判别器的输入来自真实数据分布，则训练标签为 1，如果输入样本为 $G(z)$，则训练标签为 0。

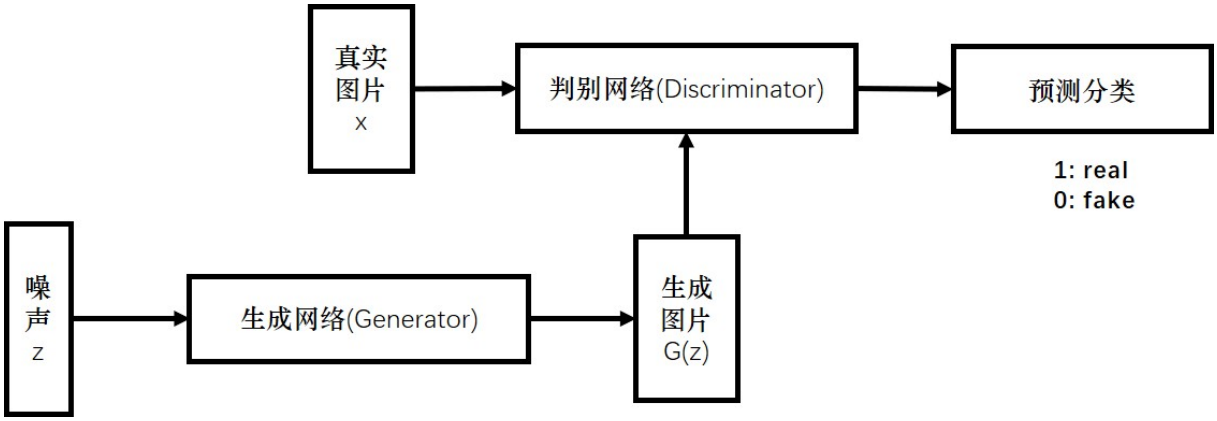


图 3.2 生成对抗网络结构图

算法 2 生成对抗网络的训练过程

输入： 生成网络模型参数 θ_G 和对抗网络模型参数 θ_D，总训练次数 n，判别网络循环次数 k，批大小 m

输出： 优化后的模型参数 $\hat{\theta_G}$ 和 $\hat{\theta_D}$

for 1:n **do**

  for 1:k **do**

    从噪声先验 $p_g(z)$ 中抽取 m 个噪声样 $\{z^1, \cdots, z^m\}$

    从数据分布 $p_data(x)$ 中也抽取 m 个真实样本 $\{x^1, \cdots, x^m\}$

    对判别器参数 θ_D 进行更新，其随机梯度表示为：

    $\nabla_{\theta_d} \frac{1}{m} \sum_{i=1}^{m} [\log D(x^{(i)}) \log(1 - D(G(x^{(i)})))]$.

  end for

  从噪声先验 $p_g(z)$ 中抽取 m 个噪声样 $\{z^1, \cdots, z^m\}$

  对生成器参数 θ_G 进行更新，其随机梯度表示为：

  $\nabla_{\theta_g} \frac{1}{m} \sum_{i=1}^{m} \log(1 - D(G(x^{(i)})))$.

end for

生成对抗网络（GAN）是一种强大的技术，它可以无人监督地生成用于训练和生成新的图像。它们在许多数据生成任务中也被证明是非常有效的，例如新的段落生成[178]。通过使用最小-最大策略，一个神经网络从原始数据分布中连续生成更好的伪造样本，以欺骗其他网络。然后训练另一个网络以更好地区分伪造品。近些年来，GAN 已被用于风格转换，例如将图像从一个风格中转换到另一个风格（CycleGAN[156]）。这些生成的图像可以用于训练汽车在夜间或雨中驾驶，仅使用例如在晴天收集的数据。此外，通过执行转移学习技术，即使使用相对较小的数据集[179]，GAN 也是有效的。此外，他们表现出非常擅长增强数据集，例如提高输入图像的分辨率[180]。GAN 是用对抗方法来生成数据的一种模型，和其他机器学习模型相比，GAN 引人注目的地方在于给机器学习引入了对抗这一理念。GAN 的结构如图3.2所示。

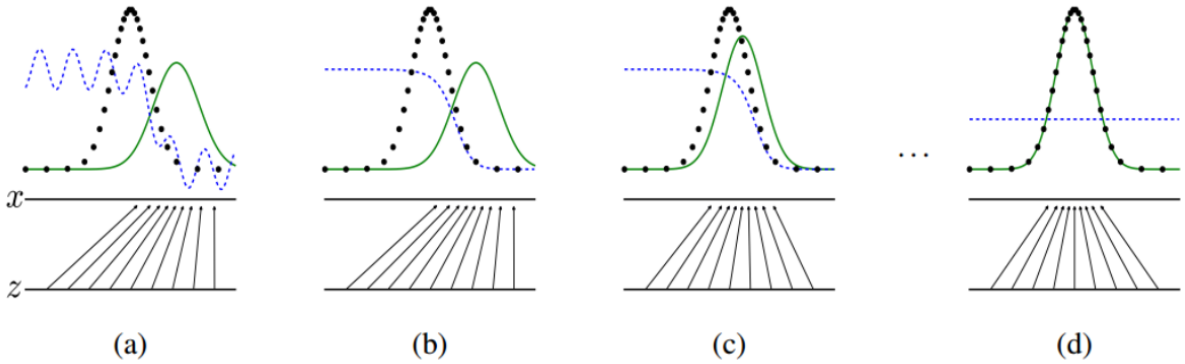


图 3.3 生成对抗网络训练过程

GAN 网络对数据分布和模型分步训练过程都总结在图3.3中。生成对抗网络通过同时更新生成网络的分布（D，表示为蓝色的虚线）来进行训练，以便区分来自数据生成分布（表示为黑色的虚线）p_x 的样本与来自生成分布 $p_g(G)$ 的样本（表示绿色的实线）。下面的水平线表示的是 z 的采样区域，在图中使用的是均匀采样。上面的水平线是 x 域的一部分。向上箭头表示的是在变换样本上施加非均匀分布 p_g 的映射 $x = G(z)$。G 在 p_g 的高密度区域样本数量多并且比较集中，而在低密度区域样本数量少并且比较分散。整个模拟的网络训练过程可以写成一下几个步骤:

(1) 展示的是一个处于快要收敛的对抗状态：p_g 和 p_{data} 已经很相似了，D 也是一个相对精确的分类器。(图3.3(a))

(2) 在算法中对 D 进行训练学习的内循环中，D 的训练目标就是用来区分生成数据和真实数据，它会收敛到 $D^*(x) = \frac{p_{data}(x)}{p_{data}(x)+p_g(x)}$。(图3.3(b))

(3) 在更新 G 之后，D 的梯度将引导 $G(z)$ 转向更可能被分类为真实数据的区域。(图3.3(c))

(4) 经过几个步骤的训练后，如果 G 和 D 有足够的表达能力，他们将达到一个平衡点，即 $p_g = p_{data}$，此时两者都无法被更新了。鉴别器不能区分两个分布，即 $D(x) = \frac{1}{2}$。(图3.3(d))

这里需要注意的是：生成模型与对抗模型是完全独立的两个模型，他们之间没有什么联系。那么训练采用的大原则是单独交替迭代训练。

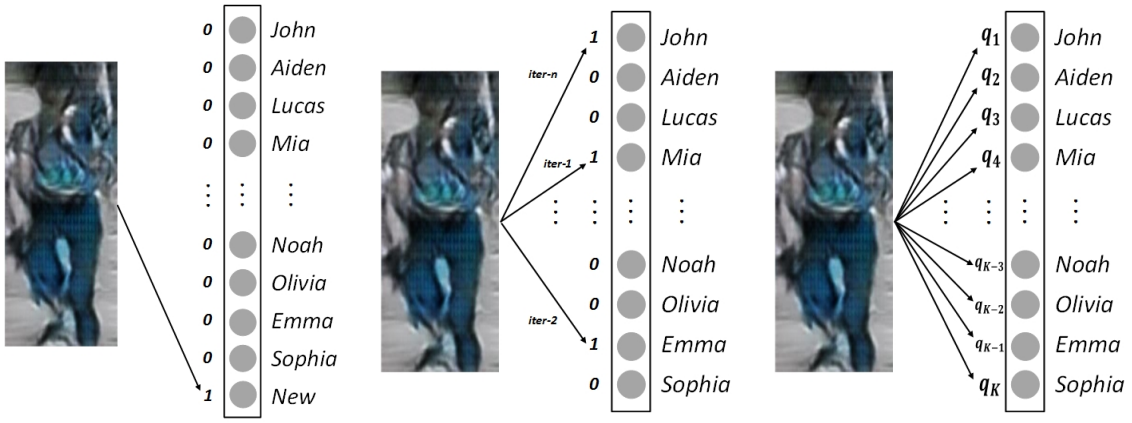
3.2.3 伪标签

手动标注由 GAN 网络生成的人物图像是不切实际的，因为对其行标注具有多方面的困难。首先，GAN 生成图像的质量并无法得十足的保证，例如生成的图片中内容的外形是否跟正常人的图片一样，另外，即使生成的图片与真实图片很接近，还是无法判断该图片属于哪个行人，因为生成图片的外观也会与现有的行人之间存在比较大的差别，而行人再识别目前主要的区别依据就在于外观。因此，为 GAN 生成的无标签行人图像的生成合理的伪标签成为了一个重要的问题。如上节所述，现在已经有一些工作开始采用 GAN 来生成样本作为行人再识别的数据增强方案并提出了一些伪标签生成方法。现有的伪标签生成方法如图3.4所示，具体的总结如下：

(1) **All-in-one**[150,151] 如图3.4(a)所示，All-in-one 寻求的是分配标签的最简单方案。它仅仅需要引入一个区分于现有的新的类，并直接将所有的无标签的生成数据分到其中，而不考虑在生成图像之间可能存在的任何差异性。在整个训练过程中的每一次迭代，这些伪标签一直使用，并不会被更新替换。

(2) **独热编码 (One-hot)**[154] 在 All-in-one 的基础上，One-hot 不仅考虑了所有生成的图像中的内部差异性，建议针对每个不同的生成样本归纳到一个现有的不同类中去。该目标类的选取是通过根据每个类的概率预测的最大值来决定的，如图3.4(b)所示。独热标签的形式与真实样本的标签形式相同，因此这种伪标签方案很容易用于网络的训练。值得注意的是，随着训练每一次的迭代，同一张生成图像的伪标签可能会有所变化，这是因为当次迭代状态的类别概率预测可能会发生变化。

(3) **分布式编码 (Distributed)**[116,152] 这种类型的伪标签是独热编码方案的进一步扩展，它的考虑基于认定无标签的数据的标签应属于分布状，如图3.4(c)中 q_i 所示。在 GAN 生成数据作为网络数据增广的前提下，由于 GAN 生成的图像是从真实数据流形中提取的伪样本，因此将它们分类为任何单个类都是不准确的。基于这个假设，Zheng 等人[116] 提出在 LSRO 的基础上分配均匀分布的标签，即 $q_i = \frac{1}{K}, \quad i = 1, 2, \cdots, K$。与之不同的是，Huang 等人[152] 则提议根据类别预测的先后顺序分配标签，来综合考虑各个现有类别对于标签的贡献。分布式伪标签与真实标签一起都可以用交叉熵损失进行训模型的学习。



(a) All-in-one (b) One-hot (c) Distributed

图 3.4 现有的三种伪标签生成方法

GAN 的一个独特的能力是它可以从训练数据分布中生成样本而无需对其进行严格地建模。换句话说，GAN 生成的图像可以被视为是从有标签样本集的数据分布中抽取出来的。因此，根据有标签样本对生成的图像进行伪标签更符合实际情况。然而，现有的伪标签生成方法都没有考虑到带标签和无标签数据之间的固有关系，用以改善半监督框架下的特征表示学习。相反，本章的工作旨在解决这一局限，通过将无标签数据与有

标签数据样本进行关联，使得网络可以自动发现这些潜在的数据分布规律，在此基础上还提出一种新的损失函数。

3.3 伪标签正则化深度模型的行人再识别

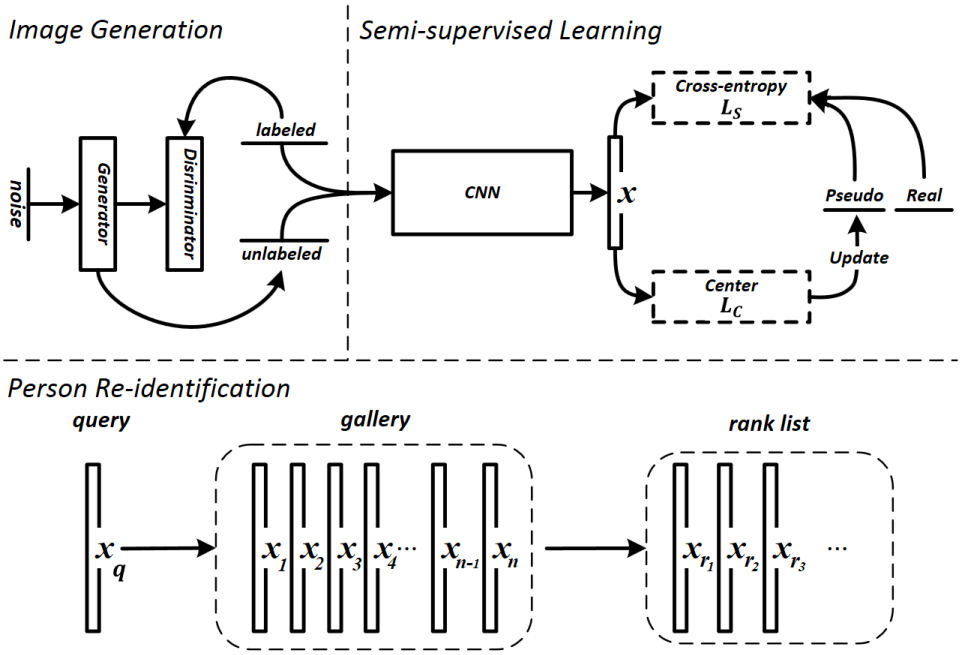


图 3.5 半监督行人再识别总体结构图

尽管来自 GAN 的生成数据样本是无标签的，但它们可以与真实样本数据一起用来训练网络，来提升特征表示的性能。这种有标签和无标签的数据一起进行特征学习的过程就是半监督学习。在这样的半监督学习的设定下，本章提出在为这些数据生成伪标签的时候，应该考虑到有标签数据中的潜在的数据分布规则，并利用这些分布方式来未无标签的数据生成伪标签。为此，本章为半监督特征学习引入了一个新的多任务学习目标函数以及两种新的伪标签编码方案。该多任务目标函数包括正常分类损失函数以及一个中心正则化项。分类损失旨在为每个行人学习身份鉴别嵌入（Identity Discriminative Embedding, IDE）[14]。中心正则化项则是为了提高了特征表示的判别能力，同时其还可以为无标签的样本提供伪标签。整个网络结构如图3.5所示。图中上排表示利用 GAN 生成合成图像的半监督学习的训练过程。该训练过程主要由两个模块组成。在左边的是第一个模块，图像生成模块。在图像生成模块中，生成对抗网络被用来利用有标签数据作为输入，对抗训练优化生成模型来估计数据分布，以便在训练完成之后，该生成器能被用来进行无标签数据的生成。生成器被训练来产生通过鉴别器测试的样本，同时训练鉴别器以将假样本与真实样本分开。最后，训练的生成器被用来生成大量的图像样本，这些样本是无标签的并是从真实数据分布的空间中抽样形成的。这些无标签数据将用于随后的训练。右边第二个模块是半监督学习部分。该模块的输入数据包含从左边网络生成

的无标签数据样本和有标签的真实数据样本，通过联合正常分类损失函数和中心正则化项来学习特征表示。下一行展示的是测试阶段的流程，所有的样本经过卷积神经网络（CNN）进行特征提取，并根据图像间的欧几里德距离来进行检索操作。

本章所提出的目标函数（如图3.5所示）可以表示为:

$$\mathcal{L} = \mathcal{L}_S + \lambda \mathcal{L}_C, \tag{3.2}$$

其中 $\mathcal{L}_S$ 和 $\mathcal{L}_C$ 分别表示分类损失和中心损失，λ 是平衡两个分量贡献的权衡参数。

3.3.1 分类损失

传统的全监督分类训练需要的是图像标签对，从 GAN 模型生成的数据则没有可用的标签。为了使用这些生成的数据进行训练，本章提出了两种方案来为无标签数据提供伪标签。后续的实证结果表明，两种方法都能够提高行人再识别的性能。首先将给出一些背景知识和定义一些符号，然后在接下来的部分将详细说明本章所提出的方法。

对于单个输入图像，卷积神经网络计算其特征表示 x 并输出其属于第 k 类的输出 y_k，其中 $y = W^T X + b$。因此，可以通过以下方式给出其属于第为 k 类的 softmax 概率:

$$p(y_k) = \frac{e^{y_k - y_{max}}}{\sum_{j=1}^{K} e^{y_j - y_{max}}}, \quad s.t., \, k \in [1, K], \tag{3.3}$$

其中 y_{max} 表示 y 中的最大响应，K 是预定义类的数量，即行人再识别数据库中行人的个数。

独热标签 一个简单的伪标签分配策略是分配一个与真实标签格式相同的伪标签，如图3.4(b)所示。参考聚类中的聚类标准，并考虑 GAN 生成和实际图像表示之间的相似性，本章采用了一种直观的解决方案，那就是将无标签的数据样本与之最相似的类进行关联。

在特征空间中一个输入数据表示 x 与第 k 类的中心 c_k 之间的相似度度量公式如下:

$$sim(\mathbf{x}, \mathbf{c}_k) = \frac{\mathbf{x} \cdot \mathbf{c}_k}{\|\mathbf{x}\| \|\mathbf{c}_k\|} \tag{3.4}$$

其中 c_k 是特征空间中 k 类的特征表示，该表示是可以动态更新的，其详细介绍可以在3.3.2中找到。无标签数据特征 x 的伪标签 ℓ 可以根据上述相似性度量来定义:

$$\ell = \arg\max_k \; sim(\mathbf{x}, \mathbf{c}_k) \tag{3.5}$$

独热伪标签的一个优点是它与真实样本标签的格式一致，这使得使用独热伪标签的无标签的数据能够与有标签数据一起送到卷积神经网络中进行训练，而无需单独的训练程序或设计新的损失函数。它们可以通过以下分类丢失函数进行训练:

$$\begin{aligned} \mathcal{L}_S &= -\log(p(y_\ell)) \\ &= -(y_\ell - y_{max}) + \log(\sum_{j=1}^{K} e^{y_j - y_{max}}) \end{aligned} \tag{3.6}$$

其回传梯度可以写为：

$$\mathcal{L}_S^{'} = \frac{e^{y_\ell - y_{max}}}{\sum_{j=1}^{K} e^{y_j - y_{max}}} - 1. \tag{3.7}$$

分布式标签 由 GAN 生成的合成图像是从逼近真实数据的流形空间中抽取的随机样本。由于高维视觉数据的复杂性，GAN 产生的行人样本可能具有模糊或荒谬的外观和体形。这些生成的低质量图像虽然可以成功通过鉴别网络的训练，但是从人的角度来检查时，它们与真实样还是很容易区分的。因此，对于最佳学习过程来说，武断地认为这些生成的无标签图像属于现有任一身份，并且按照上面所述分配一个独热标签是不合适的。为此，将单个无标签数据近似地视为来自不同类别的行人表示的加权组合是一个相对来说更优的选择。该方案在图3.4(c)中示出。最终分布式标签 $q(y)$ 由 x 和所有聚类中心 c 之间的相似性的 softmax 函数定义，公式如下：

$$q(y_k) = \frac{e^{sim(\mathbf{x},\mathbf{c}_k)}}{\sum_{j=1}^{K} e^{sim(\mathbf{x},\mathbf{c}_j)}}, \tag{3.8}$$

因此，分布式伪标签可以被解释无标签数据属于每个已知类的概率。正如文献[153]中所建议的，交叉熵函数可用于训练具有分布式编码的伪标签，对于单个输入，其分类损失计算如下：

$$\begin{aligned} \mathcal{L}_S &= -\sum_{k=1}^{K} q(y_k)\log(p(y_k)) \\ &= -\sum_{k=1}^{K} q(y_k)(y_k - y_{max}) + q(y_k)\log(\sum_{j=1}^{K} e^{y_j - y_{max}}) \end{aligned} \tag{3.9}$$

相应地，回传梯度可以写为：

$$\mathcal{L}_S^{'} = q(y_k)\frac{e^{y_k - y_{max}}}{\sum_{j=1}^{K} e^{y_j - y_{max}}} - q(y_k) \tag{3.10}$$

具体来说，如果针对 $q(y_k)$ 进行约束，使其满足

$$q(y_k) = \begin{cases} 1 & k = \ell, \\ 0 & k \neq \ell. \end{cases}$$

并将其带入公式(3.8)和公式(3.10)，将可以分别获得(3.6)和公式(3.7)。这也就意味着分布式伪标签方案是独热伪标签的一种推广形式，并且这两种由特征相似度得出的伪标签都可以在一个统一的架构中利用原始的交叉熵损失函数的进行训练。

3.3.2 中心损失

除了先前的分类损失，本章还提出了为分类损失增加了一项中心正则化项[181]。该项能够将特征空间中无标签样本与其相匹配聚类中心进行关联。它通过基于中心的聚类发现特征空间中的隐藏的数据分布模式，从而执行智能的数据扩充。

给定一个特征空间中批大小为 m 的图像特征表示 $\{x_i \in \mathbb{R}^d\}_1^m$，中心正则化损失可以写成以下形式：

$$\mathcal{L}_C = \frac{1}{2}\sum_{i=1}^{m}\|\mathbf{x}_i - \mathbf{c}_{\ell_i}\|_2^2 \tag{3.11}$$

其中 $c_{l_i} \in \mathbb{R}^d$ 是 x_i 所属的簇的中心特征，d 为特征表示的维度。对于正在处理的是具有分布式伪标签的数据的情况，c_{l_i} 的选择是根据公式(3.5)进行选择的，这样的选择是为了训练过程的简单性。该中心正则化项一方面考虑了类内特征的完备性，可以与分类损失函数联合起来获得一个更加鲁棒的行人特征表示；另一方面在计算过程中可以获得所有类别的中心，这有助于为无标签数据生成伪标签，同时，这样也能减少每次类中心特征的计算量。

当前向损失计算完成后，可以通过计算下面公式来得到对于 x_i 的回传梯度：

$$\mathcal{L}'_C = \mathbf{x}_i - \mathbf{c}_{\ell_i} \tag{3.12}$$

接下来，使用以下方法更新集群中心特征表示：

$$\Delta\mathbf{c}_k = \frac{\sum_{i=1}^{m}\delta(\mathbf{x}_i \in \mathbf{x}^L)\cdot\delta(\ell_i = k)\cdot(\mathbf{c}_k - \mathbf{x}_i)}{1 + \sum_{i=1}^{m}\delta(\mathbf{x}_i \in \mathbf{x}^L)\cdot\delta(\ell_i = k)} \tag{3.13}$$

其中 X^L 是有标签训练数据的集合，δ 表示 delta 函数，即 $\delta(\cdot) = 1$，如果条件满足，否则为 0。

值得注意的是相较于原版的中心损失，网络的训练过程中有三个值得注意的细节：

(1) 类中心特征的更新是基于每一个小批量的样本，而不是基于整个训练集，减少了计算复杂度。

(2) 仅仅用批中包含的有标签数据来对类中心特征进行更新，而不使用无标签的数据，这样有助于类中心特征形成过程的稳定性，因为无标签数据的类别会出现很大的波动。

(3) 正则项的损失计算是所有有标签和无标签样本都参与的，但其回传的梯度仅仅适用于每个批数据中有标签的样本中。

对于一批输入样本以及两种不同编码形式（包含独热和分布式）的伪标签，其最终的损失函数(3.2)可以统一地写成下面的形式：

$$\begin{aligned}\mathcal{L} &= \mathcal{L}_S + \lambda\mathcal{L}_C \\ &= -\sum_{i=1}^{m}\sum_{k=1}^{K} q(y_k^i)\log(p(y_k^i)) + \frac{\lambda}{2}\sum_{i=1}^{m}\|\mathbf{x}_i - \mathbf{c}_{\ell_i}\|_2^2\end{aligned} \tag{3.14}$$

其中 y_k^i 表示第 i 个输入图像的 y_k。

本章所提出的基于特征相似性的半监督特征学习过程可以表示为算法3。

3.3.3 讨论

在本节中，将研究和展示本章所提出的方法的一些有趣的特性，并将其与现有工作进行比较。

算法 3 基于伪标签的半监督特征学习方法

输入: 带标签数据 L，生成的无标签数据 U，最大迭代次数 T，批尺寸 m，中心更新率 α，权重参数 λ，网络参数 θ

输出: 优化后的模型参数 $\hat{\theta}$

初始化: 训练集 $X = L \cup U$, 利用预训练的 ResNet-50 来初始化模型参数 θ, $\alpha = 0.5$, $\lambda = 10^{-4}$, 类中心点 $\{\mathbf{c}_k = \mathbf{0} | k = 1, 2, ..., K\}$

for $t = 1 : T$ **do**

重新打乱 X 排序并从中取样 m 个样本得到一个小的训练数据批 X^t,

将训练数据批 X^t 通过网络向前传播从而得到他们的特征表示 $\mathbf{x}^t$,

for $\mathbf{x}_i^t$ 属于无标签的数据 U **do**

利用公式(3.4)计算出每个特征与每个类别中心点之间的相似度 $sim(\mathbf{x}_i^t, \mathbf{c}_k)$,

分别利用公式(3.5)和(3.8)可以用来为每个表示 $\mathbf{x}_i^t$ 生成独热码和分布码类型的伪标签 $\ell_{\mathbf{x}_i^t}$,

end for

利用 $\mathbf{x}^t$ 和公式(3.14)来计算出最终的组合损失 $\mathcal{L}^t$,

if $\mathbf{x}_i^t$ 属于有标签的数据 L **then**

利用公式(3.13)计算类别中心点更新量 $\mathbf{c}^t$,

利用 EMA 更新类别中心点表示 $\mathbf{c}^t = \mathbf{c}^{t-1} + \alpha \Delta \mathbf{c}^t$,

end if

梯度回传,

对模型参数 θ^t 进行更新

end for

返回: $\hat{\theta} = \theta^T$

特征相似性 vs 类别预测 本章工作与之前关于伪标签生成的所有工作之间的一个显著差异是，本章的工作是第一个提出基于特征空间中的特征表示相似性来进行伪标签的生成的。一种基于深度学习的常见行人再识别方法的工作原理是，他们首先训练一个身份分类网络，然后提取最后连接的层激活作为最终描述符，以便在后续测试阶段进行相似度计算。诸如文献[152,154]的先前工作是基于分类预测概率来计算伪标签。如果使用具有最大概率响应的类用作训练标签，则网络预测的方法可以直接用于独热伪标签的生成[154]。然而，在分布式标签的情况下这种概率就会失效，因为伪标签将与类概率的预测相同，这将会导致最终不会产生任何权重校正，从而无法进行反向传播。因此，Huang 等人[152] 提出对预测概率进行排序并基于该排序先后分配分布式的伪标签。这

样的排序会不可避免地引入了不准确之处。相反，本章则建议将伪标签生成本身视为与最终再试别相类似的检索过程，将其中无标签数据作为查询，有标签数据作为图库，利用特征表示的相似性进行检索，这与行人再识别中进行的最终检索模式相同。以这种方式，基于相似性的伪标签生成方案适用于独热编码和分布式编码标签的生成。

为什么中心重要？ 该中心正则化项的引入起到了非常重要的作用，其对于半监督设定下的行人再识别的贡献主要有以下几点：

(1) 它能够帮助卷积神经网络学习到更加鲁棒的特征表示。传统的分类损失函数主要在于区分开类间的差异，而并不考虑类内的紧凑性，使用了中心正则化项之后就能够同时考虑到类内的紧凑性，从而帮助网络学习到更加具有鉴别性的行人特征表示[181]。

(2) 它能够为无标签的数据生成伪标签。由于伪标签的计算是基于无标签数据的特征与各类中心特征之间的特征相似度计算出来了，于是在有了它的情况下就能够很方便地计算出相似度从而生成伪标签。

(3) 它能够减少神经网络的计算量。该类中心特征采用的是动态更新方法，这样的动态中心特征更新方法能够在网络的训练过程中就计算出类中心特征，而不需要将所有的特征全部提取出来再计算，大大减轻了神经网络的计算量。

基于中心的方法考虑到了无标签与有标签数据之间的关系，并且依据这样的关系来为这些无标签的数据产生伪标签。这是一个很直观的考虑，因为可以假设一个类的生成样本接近原始样本应该是基于特征空间中它们之间的相互关系。相反，以前的一些工作如文献[154]和文献[152]不恰当地用预测概率来作为数据伪标签的生成标准，并没有考虑它们与有标签数据的固有关系。

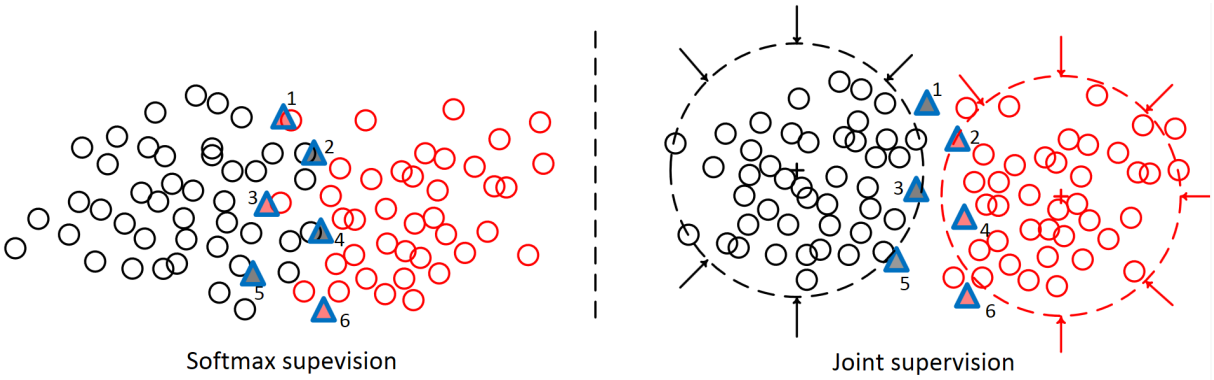


图 3.6 基于预测和相似度的标签分配对比

为了能够直观的展现出基于中心特征相似性与基于预测的方法的差异性，图3.6中展示了一个示例来可视化出两种不同的伪标签分配结果的差异性。图3.6中的左侧部分表示的是仅使用了 softmax 损失来进行的二分类问题（黑色和红色）数据集上的分类。同时还有一些无标签的数据样本（显示为蓝色三角形）散布在两个类之间的边界周围。如果使用像 Lee 等人[154] 提出的基于概率的伪标签方法，那么其伪标签的寻找过程可

以粗略地看作是最近邻搜索。样本 1 和样本 3 在特征空间中的最近邻应该都是红色的样本，因此，它们的伪标签将被设置为红色。然而，实际情况是，在特征空间中，从一个比较大的范围来看，它们更倾向于黑色。类似地，样本 2 和样本 4 会被错误地分类为黑色尽管它们应该更倾向于红色。图3.6中右边的部分是加入中心损失后的标签结果。值得注意的是，在这样的情况下，红黑两类本身就已经被分开的比较好，再基于无标签数据特征到两类的中心特征的相似性来决定其伪标签而不是通过与其最近的样本点的标签来决定，这样的伪标签生成方法更加合理。

表 3.1 多种伪标签生成算的对比

方法	标签分配	标签编码	标签来源	各类贡献
All-in-one[150,151]	静态	独热	手工	-
One-hot[154]	动态	独热	概率	-
LSRO[116]	静态	分布	手工	相同
MpRL[152]	动态	分布	概率	不同
FAPL-o	动态	独热	相似性	-
FAPL-d	动态	分布	相似性	不同

与相近工作的比较 表3.1总结了与本章所提出的方法密切相关的方法的一些属性的总体比较。在接下来的部分种，本章提出的两种方案将分别表示为 FAPL-o（独热）和 FAPL-d（分布式）。目前现有的应用于行人再识别对 GAN 生成数据进行伪标签生成的策略包括 All-in-one[150,151]，One-hot[154]，LSRO[116] 和 dMpRL[152]。它们的标签分布可以分别如图3.4(a)，图3.4(b)和图3.4(c)所示。其中，LSRO 和 dMpRL 都采用了分布式编码的伪标签，但其两者的不同之处在于 LSRO 选择均匀的标签分布，即该伪标签的分布是一个均匀分布，而 dMpRL 则是基于类别预测概率的排名来生成伪标签的。

与 Odena 等人[150] 和 Zheng 等人[116] 提出的直接为所有生成的无标签数据分配固定和相同的标签相比，本章所提出的方法考虑了它们与有标签数据之间的关系，并随着训练的进展动态预测每次迭代中的伪标签。Huang 等人[152] 和 Lee 等人[154] 都提出使用动态的标签分配机制，但两者都采用类概率预测而不是特征相似性来分配标签。本章工作则利用的是特征空间中的特征之间的相似性来相应地预测标签。并且实验结果表明，与固定的独热标签相比，分布式概率标签更具弹性和灵活性。总而言之，本章所提出的方法具有更灵活，更具鉴别性并且能够识挖掘特征空间中更大范围潜在数据规律的优点。

3.4 实验结果与分析

在本节实验部分中，将在三个广泛采用的行人再识别数据集上进行了实验，来对本章所提出的方法的有效性进行评价。

3.4.1 实现细节

性能基准 在本章的实验中，采用的是 He 等人[124] 提出的标准 ResNet-50 作为骨干架构。原因在于该网络架构已用于评估密切相关的伪标签方法，例如 All-in-one[150,151]，One-hot[154]，LSRO[116] 和 dMpRL[152]。用于训练的网络结构并没有发生任何其他的改变，除了在最后一层把 1000 类激活神经元替换行人数据库中行人 ID 数量，即 Market-1501 和 DukeMTMC-reID 的 751 和 702。在训练过程中，首先将所有训练图像调整为 256×256，然后随机水平翻转并裁剪为输入尺寸 224×224。在最终卷积层之前插入具有 0.75 丢弃率的丢失层，以防止该网络对数据集产生过拟合的现象。整个模型由随机梯度下降（SGD）对参数进行优化，动量设置为 0.9。学习率在前 40 个时期设定为 0.001，在剩下的 10 个时期衰减到 0.0001。在测试期间，提取最后全连接层 2048 维的激活作为行人描述符。该描述符会在之后进行基于余弦相似性的排序。整个网络的训练使用的是 MatConvnet[135] 的深度学习框架。

GAN 模型 为了公平比较，实验的设定都遵循文献[116]中的训练标准，并同样地采用 DCGAN[157] 作为生成模型来生成假的无标签的行人图像，用于训练集的扩大从而对模型进行正则化。该 DCGAN 的生成器的输入是由一个 100 维的随机向量构成，通过一个线性函数将其放大以形成 $4 \times 4 \times 16$ 的张量，随后总共通过 6 个反卷积层 (deconvolution layers) 和 5×5 大小的卷积核 (kernel) 以获得期望的大小为 $128 \times 128 \times 3$ 的图像。鉴别器中包含了 5 个卷积层，每个卷积层的卷积核大小都为 5×5，通过它来执行二分类任务以从给定输入图像中分离真实的和伪造样本。在网络训练完成之后，利用训练好的生成器随机产生了 36,000 张合成图像。对于以下半监督学习，将所有合成数据样本的大小都调整为 256×256。一些生成的数据样本在图3.7中做了展示，尽管一些生成的图像和真实图片相差甚远，但实验证明它们仍然有助于模型的正则化并提高性能。

特征提取 在特征提取的过程中，特征的时间消耗也是一个值得关注的点。如上所述，该网络结构使用 ResNet-50 作为骨干框架，并采用本章提出的新的损失函数进行特征的学习。因此，对于任意一张输入图像，其特征提取的过程仅仅为一次 ResNet-50 进行前向传播所需要的时间。在实验中，在一张 NVIDIA TITAN Xp 卡上对单幅图像进行特征提取时，仅采用前向传播的特征提取大约需要 40 毫秒。值得注意的是，当使用一个批大小为 100 张的图像输入时，单个图像的平均提取时间会被进一步减少到 3 毫秒。因此，本章的方法在特征提取阶段相对于基础网络而言并没有增加任何的计算量。

3.4.2 性能评估

方法的有效性 该方法的整体实验结果总结在表3.2中。如表中所示，当使用 ResNet-50 作为主干架构，在 Market-1501，DukeMTMC-reID 和 CUHK03 数据集上的基准性能分别达到 72.74%，65.22% 和 70.68% 的 rank-1 准确度以及 50.99%，44.99% 和 74.25% 的 mAP。在这个方法有效性实验中，通过随机选择了三个数据集上生成的 24,000 张图

图 3.7 不同 GAN 生成图像之间的比较

像作为无标签的数据进行辅助训练，观察到本章所提的两个方案比基准性能都有了明显的提升。例如，在 Market-1501 数据集上，FAPL-o 取得了 82.04% 的 rank-1 准确率，比基准性能提高了大约 10%，同样地，mAP 值则是从 50.99% 提升到了 61.26%。FAPL-d 更是在 FAPL-o 的基础上进一步地提升了性能。这样的性能增长也可以在 DukeMTMC-reID 和 CUHK03 数据集上看到。在这两个数据集上，平均的 rank-1 准确率增长为 4%。此外，还可以观察到的是在 DukeMTMC-reID 上的性能增长相对要低于 Market-1501，这样的现象的原因是由于 DukeMTMC-reID 数据集中存在着严重的遮挡而导致该数据集更加具有挑战性。

表 3.2 与伪标签生成算法性能比较

方法	Market-1501		DukeMTMC-reID		CUHK03	
	rank-1	mAP	rank-1	mAP	rank-1	mAP
Baseline	72.74	50.99	65.22	44.99	70.68	74.25
LSRO[116]	78.21	56.33	67.68	47.13	73.10	77.40
dMpRL-II[152]	80.37	58.59	68.24	48.58	68.68	73.48
FAPL-o (Ours)	82.04	61.26	70.92	51.99	73.28	78.92
FAPL-d (Ours)	**83.43**	**63.23**	**71.90**	**52.25**	**74.17**	**79.62**

无标签数据量 在本节，将通过控制使用不同数量的无标签数据参与训练来评估其对于行人再识别性能的影响。如前面所说，在实验中，首先使用 DCGAN 生成了 36,000 张图像，然后从中随机挑选组合子集以来进行评估。其结果展示在表3.3中。首先看到的

是，对于本章所提出的两种标签编码方案，随着无标签数据的加入，模型的性能得到了显著的提升，两个数据库的 rank-1 准确率在基准的基础上分别提升了 9.38% 和 10.69%，mAP 值则是在基准的基础上提升了 11.32% 和 12.24%。然而，当无标签图像的数量增加超过一定的阈值之后（即：从 12000 到 36000），其性能并未能进一步稳定地提升，而是处在了一种微微波动的状态，对于 FAPL-o 来说，其波动在 82% 左右而 FAPL-d 则波动在 83% 左右。这个现象并不仅仅存在于本章所提出的方法，而是在所有其他伪标签方法中普遍存在的。经过推测，这种现象的原因是由于生成图像的表达能力有限而导致的。因为这些无标签的图像是从特定的流型分布（真实数据的流形）中采样的，因此简单地增加样本数量不会为模型提供任何有益于最终检索任务的额外信息。

表 3.3 不同数量生成图像对性能的影响

GAN 图像数	All-in-one[150,151]		One-hot[154]		LSRO[116]		sMpRL[152]	
	rank-1	mAP	rank-1	mAP	rank-1	mAP	rank-1	mAP
0 (基准)	72.74	50.99	72.74	50.99	72.74	50.99	72.74	50.99
12000	76.96	55.68	76.52	55.69	77.17	55.22	77.73	55.27
18000	**77.40**	55.59	77.95	55.04	76.96	55.28	77.73	55.05
24000	77.21	56.07	77.62	**56.90**	**78.21**	**56.33**	**78.85**	55.59
30000	77.17	**56.19**	**77.95**	56.54	77.46	55.40	77.82	**55.76**
36000	75.92	55.24	77.42	56.38	77.91	55.82	78.32	55.45
性能提升	4.66	5.20	5.21	5.91	5.47	5.34	6.11	4.77

GAN 图像数	dMpRL-I[152]		dMpRL-II[152]		FAPL-o(Ours)		FAPL-d(Ours)	
	rank-1	mAP	rank-1	mAP	rank-1	mAP	rank-1	mAP
0 (基准)	72.74	50.99	72.74	50.99	72.74	50.99	72.74	50.99
12000	77.88	55.84	79.22	58.14	81.38	60.31	**83.28**	**61.68**
18000	78.36	56.21	79.81	58.31	82.10	**62.31**	**83.16**	**62.38**
24000	77.79	56.10	**80.37**	**58.59**	82.04	61.26	**83.43**	**63.23**
30000	78.65	57.15	79.16	57.69	82.10	61.42	**83.02**	**62.41**
36000	**78.95**	**57.42**	79.90	57.61	**82.12**	60.70	**82.30**	**61.92**
性能提升	6.21	6.43	7.63	7.60	9.38	11.32	**10.69**	**12.24**

有标签数据量 在学习任务中，能够用越少的数据来进行鲁棒的特征表示是非常受欢迎的。为了能够测试无标签数据在有标签数据量非常有限的情况下是否对于性能还能够有所提升和帮助，本节在减少的训练集上进行实验。在实验中，将随机将训练集有标签的数据的数量大致减少到总量的一半和三分之一来模拟一种比较极端的情况。在生成

表 3.4 Market-1501 数据集更少标注数据实验结果

GAN 图像数	All		half		third	
	rank-1	mAP	rank-1	mAP	rank-1	mAP
0(基准)	72.74	50.99	66.98	43.71	57.45	33.22
12000	83.28	61.68	76.81	54.25	66.39	42.87
18000	82.16	61.68	77.02	54.48	65.23	41.22
24000	**83.43**	62.23	77.46	55.26	66.48	42.04
30000	83.02	**62.41**	**78.59**	**56.50**	66.69	43.50
36000	82.30	61.92	78.15	56.30	**67.87**	**43.54**

这些子集时，遵循了以下的筛选规则：（a）如果一个行人身份只有少于 8 张的行人图像，则将会保留其所有样本;（b）对于超过 8 张样本的行人身份，将仅仅保留其半数或三分之一的样本。最终，可以获得了具有 7,106 个训练样本的 half 子集和具有 4,200 个训练样本的 third 子集。实验的结果可以在表3.4中找到。首先可以看到的是，当有标签的数据量逐步减少之后，需要更多无标签的数据辅以训练才能获得最佳的性能。此外，随着有标签数据减少到一半和三分之一，基准模型的性能逐步下降，rank-1 准确率分别从 72.74% 降至 66.98% 和 57.45%。这结果是符合预期的因为减少了有标签的样本量，该训练集所能提供的监督信息变少，所以导致了性能的下降。然而，通过采用本章提出的方法，可以观察到所有行的 rank-1 准确率和 mAP 值相对于基准值都有了大约 10% 的提升。具体而言，通过引入无标签的数据，half 模型（rank-1=78.59%，mAP=56.5%）成功地超过完全监督的基准模型的性能（rank-1=72.74%，mAP=50.99%），并且有着大约 5% 的提升。另外一个值得注意的点是，在 third 数据集上，参与训练的无标签数据有 36,000 张图像，大约是有标签数据 (4,200 张）的 8.5 倍，该的模型仍然去的了最佳的性能表现（rank-1=67.87%，mAP=43.54%），这是结果正表明本章的方法可以应用到更小的数据集上来进行半监督的学习。

表 3.5 权重参数 λ 的实验结果

λ	独热编码		分布式编码	
	rank-1	mAP	rank-1	mAP
0.001	80.82	60.42	81.05	61.18
0.0001	**82.04**	**61.26**	**83.43**	**63.23**
0.00001	81.18	59.42	82.31	62.31

参数灵敏度 为了探究公式(3.2)中中心正则化项的权重大小对于网络最终性能的影响，在本节中进行了参数 λ 的灵敏度测试，其结果如表3.5所示。在本章的实验中，λ

默认设置为 10^{-4}。在此之外，本章尝试设置此参数为不同的值（例如 10^{-3} 和 10^{-5}）来观察其对性能的影响，可以注意到的是在上述两种情况下，FAPL-o 和 FAPL-d 的性能均有所下降。对于 FAPL-o 来说，rank-1 准确度分别从 82.04% 下降到 80.82% 和 81.18%，mAP 从 61.26% 下降到 60.24% 和 59.42%。FAPL-d 的性能变化趋势也与此类似。根据这样的性能变化趋势来推测，将 λ 设置成 10^{-4} 能够取得比较好的效果的原因在于这样的设置会使得两个损失项在同一个数量级上，从而具有可比性。

表 3.6 DCGAN 与 IWGAN 生成图像在 Market-1501 数据集实验结果

GAN 图像数	Market-1501							
	DCGAN				IWGAN			
	One-hot		Distributed		One-hot		Distributed	
	rank-1	mAP	rank-1	mAP	rank-1	mAP	rank-1	mAP
12000	81.38	60.31	**83.28**	61.68	82.84	62.87	83.05	**62.60**
18000	82.10	62.31	82.16	61.18	82.17	62.50	**82.89**	**62.26**
24000	82.04	61.26	83.43	63.23	82.42	61.58	**83.84**	**63.41**
30000	82.10	61.42	83.02	62.41	82.66	61.62	**83.58**	**63.78**
36000	82.12	60.70	82.30	61.92	82.51	61.29	**82.78**	**63.43**
GAN 图像数	DukeMTMC-reID							
	DCGAN				IWGAN			
	One-hot		Distributed		One-hot		Distributed	
	rank-1	mAP	rank-1	mAP	rank-1	mAP	rank-1	mAP
12000	71.57	52.68	71.68	52.83	72.21	53.23	**73.16**	**54.96**
18000	70.38	51.87	71.32	52.88	71.98	53.34	**72.60**	**53.97**
24000	70.92	51.99	71.90	52.25	71.72	53.01	**72.35**	**54.00**
30000	70.47	52.54	72.38	53.71	72.48	53.35	**73.44**	**54.71**
36000	71.23	52.18	72.40	53.73	72.40	52.76	**72.85**	**53.85**

不同质量的无标签数据 为了更好地发现生成的不同质量无标签图像对于模型的正规化的效果，本章还挑选了最近被提出来的新的一个 GAN 模型来实现无标签图像生成。在这里，本节选择的是最近提出的改进后的 Wasserstein GAN（IWGAN）[159] 用于图像生成。与 DCGAN 相比，IWGAN 具有更强大的理论保证，因为其使用的是 Wasserstein 距离测量作为对抗性训练损失，该损失函数的选择可以提供更快训练速度和更稳定的收敛性。

对于生成器，首先生成 128 维的随机噪声向量并使用 5 个 3×3 残差（使用 skip-layer）

的反卷积层 (deconvolutional layers) 对其进行上采样以获得一个尺寸为 $128\times128\times4$ 的特征图，然后该特征图再经过另一个 3×3 的卷积层以生成最终大小为 $128\times128\times3$ 的输出样本。鉴别器将 $128\times128\times3$ 图像作为输入，并首先通过卷积层以获得 $128\times128\times64$ 的中间表示，然后通过另外 5 个残渣下采样卷积层得到一个 8192 维的表示，然后采用与 DCGAN 一样的二类分类器来预测输入是真实的还是假的。与 DCGAN 图像生成类似，输出图像的大小调整为 256×256，以用于本章的模型训练。

表 3.7 CUHK03 数据集实验结果

方法	CUHK03	
	rank-1	mAP
Gate-reID[80]	68.10	58.84
LOMO+XQDA[12]	46.30	-
dMpRL-II[152]	68.68	73.48
LSRO[116]	73.10	77.40
SVDNet[147]	**81.80**	**84.80**
Baseline	70.68	74.25
FAPL-o+DCGAN	73.28	78.92
FAPL-d+DCGAN	74.17	79.62
FAPL-o+IWGAN	73.99	78.64
FAPL-d+IWGAN	74.58	79.89
FAPL-d+IWGAN+re-rank	**80.73**	**86.38**

两个 GAN 网络生成的样本在图3.7中进行了比较。可以注意到，不管是 DCGAN 还是 IWGAN，它们生成的图像与真实图像都是不可比的，并且存在着很大的差距。但同时，底行中显示的 IWGAN 图像在视觉上要优于 DCGAN(中间一行) 生成的图像。具体来讲，DCGAN 生成的图像具有比较少的多样性，反而包含相当大的扭曲以及模糊的肢体和身体形状。相反，IWGAN 可以更好地保留人体形状，并可以生成更真实的样本，这些样本的衣物颜色变化也很大。

对于两种图像生成方法，都通过随机的方式选择了不同数量的生成图像作为真实训练集的补充来进行网络的训练，并在表3.6中给出了实验结果。在 Market-1501 数据集上，相同的独热伪标签设置下，DCGAN 在 rank-1 准确率达到 81.38%，而 IWGAN 则是达到了 82.84%。相比之下，在 DukeMTMC-reID 数据集上，在分布式标签的设置中，当使用的样本个数为 12,000 时，其 rank-1 的准确率有了 1.48% 的增加。总体而言，可以从该实验得出两个结论：（a）当使用的生成图片有着更好的视觉效果和多样性的时

候，该方法在不同的数据库上的性能都能进一步提高 0.5%-1% 左右，说明了无标签数据的视觉效果对模型的正则化是有帮助的。但是，这种性能的提升相对来说比较小，产生这样的结果的原因是由于训练中采用的无标签数据的量都是很大的，通过大量的数据样本的输入减少了劣质样品带来的影响。(b) 无论采用哪种视觉质量的无标签数据来辅助训练，分布式伪标签方法的性能要始终优于独热伪标签方法，这也就说明了使用分布式的伪标签更加符合数据的属性。

表 3.8 Market-1501 数据集实验结果

方法	Market-1501		方法	Market-1501	
	rank-1	mAP		rank-1	mAP
Gate-reID[80]	65.88	39.55	SVDNet[147]	82.30	62.10
SCSP[40]	51.90	26.35	Part Aligned[182]	81.00	63.40
DNS[19]	61.02	35.68	PDC[29]	84.14	63.41
ResNet+OIM[148]	82.10	-	LSRO[116]	78.06	56.23
Latent Parts[133]	80.31	57.53	dMpRL-II[152]	80.37	58.59
P2S[183]	70.72	44.27	Baseline	72.74	50.99
re-rank[95]	77.11	63.63	FAPL-o+DCGAN	82.10	62.31
Consistent-Aware[184]	80.90	55.60	FAPL-d+DCGAN	83.43	63.23
Spindle[185]	76.90	-	FAPL-o+IWGAN	82.66	61.62
SSM[107]	82.21	**68.80**	FAPL-d+IWGAN	83.58	63.78
JLML[141]	**85.10**	65.50	FAPL-d+IWGAN+re-rank	**86.07**	**77.64**

与其他伪标签方法的比较 本节中，将针对所提出的方法在 Market-1501 数据集与现有的四种伪标签方法进行比较。比较的伪标签方法包括 All-in-one[150,151], One-hot[154], LSRO[116] 和 MpRL[152]。在所有的工作中，LSRO[116] 是目前使用均匀分布伪标签来进行模型正则化中效果最好的。而 MpRL[152] 是一项基于 LSRO 的工作，它通过考虑类别概率的贡献度来改善分布式标签，并得到了非常有竞争力的结果。Huang 等人[152] 提供了三种不同的实现方法，分别是 sMpRL，dMpRL-I 和 dMpRL-II。具体地说，第一种实现方式 sMpRL 在整个训练过程中分配了固定的分布式标签，这类似于 LSRO，除了在生成标签时考虑不同类不同的贡献而非均匀分布的标签。dMpRL-I 和 dMpRL-II 都为每个生成的样本动态分配伪标签，但两者的区别在于何时加入无标签数据用于训练。对于 dMpRL-I 来说，在训练开始的时候便引入使用生成的无标签数据，而 dMpRL-II 则是选择首先将网络训练 20 个时期之后，当 CNN 网络相对稳定时，再引入使用生成的无标签数据。表3.3还总结了其他基于伪标签的行人再识别工作的结果。可以看到的是，当参与训练的无标签数据为 24,000 张时，LSRO[4] 在 Market-1501 数据集上获得

表 3.9 DukeMTMC-reID 数据集实验结果

方法	DukeMTMC-reID	
	rank-1	mAP
BOW+kissme[8]	25.13	12.17
LOMO+XQDA[12]	30.75	17.04
LSRO[116]	67.68	47.13
dMpRL[152]	68.24	48.58
Verif + Identif[186]	68.90	49.30
APR[139]	70.69	51.88
ACRN[187]	72.58	51.96
PAN[128]	71.59	51.51
FMN[188]	74.51	56.88
Bilinear Coding[189]	76.20	56.90
SVDNet[147]	76.70	56.80
DPFL[142]	**79.20**	**60.60**
Baseline	65.22	44.99
FAPL-o+DCGAN	71.57	52.68
FAPL-d+DCGAN	72.38	53.71
FAPL-o+IWGAN	72.40	52.76
FAPL-d+IWGAN	72.85	53.85
FAPL-d+IWGAN+re-rank	**79.04**	**70.74**

了最佳性能，其中 rank-1=78.21%，mAP=56.33%。可以观察到三种 MpRL[152] 实现方法的都能够提高再识别的性能，并且 dMpRL-II 可以达到最佳结果，rank-1=80.37%，mAP=58.59%。本章提出的独热编码的伪标签方法超过了 dMpRL-II[152]，在 rank-1 准确率和 mAP 方面分别比 dMpRL-II[152] 高 1.75% 和 3.72%，分布式伪标签方案则进一步将性能提高到 3.06% 和 4.64%。这样的性能提升是合理的，因为分布式标签考虑了来自每个类的相似性贡献，并且更适合于 GAN 生成的数据。另一个值得注意的事实是，尽管本章提出的无标签策略是从训练开始时便对未无标签的数据生成伪标签并进行训练的，而后者 dMpRL-II[152] 在 20 个时期之后卷积神经网络相对稳定情况下才开始使用无标签的数据，本章提出的方法仍然取得了比他们更好的效果。

与最先进的方法比较 尽管本章的工作主要致力于研究如何更好地利用 GAN 网络生成的图像来对深度神经网络模型进行正则化，从而提升神经网络的性能，本节也将本

章提出的方法与现有的效果很好的行人再识别方法进行了比较。如表3.8中所示，使用IWGAN 生成图像的分布式伪标签方法在 Market-1501 数据集上获得了 rank-1=83.58%, mAP=63.78%, 除了 JLML[141] (rank-1=85.1%, mAP=65.50%) 和 PDC[29] 之外, 它与许多最先进的方法相比性能不相上下。JLML[141] 的 rank-1 准确率比本章的方法高 2% 的主要原因是因为本章的网络仅仅是一个单分支架构，而 JLML[141] 则使用了三个专注于不同局部区域的网络进行组合。利用文献[95]中提出的最先进的再排序技术，本章所提的方法的 rank-1 准确率可以进一步提升 2.5%, mAP 则可以进一步提升 13.86%, 这样的结果表明了该方法所学的特征中实包含了比较好的近邻结构信息的。而在 DukeMTMC-reID 数据集（表3.9）上，加上了再排序技术之后，该模型取得了 79.04% 的 rank-1 准确率和 70.74% 的 mAP。DFPL[142] 在 rank-1 准确率上略微高出（约 0.2%），因为 DFPL[142] 利用了具有不同输入规模的多个网络，并使用了一致性学习，迫使属于同一身份不同尺寸的行人表示相接近。然而，本方法的 mAP 达到 70.74%，比 DFPL 高出 10%（60.60%）。在 CUHK03 数据集（表3.7）中，可以清楚地看到，本章提出的模型在 rank-1 中分别比两个性能最近的伪标签生成方法方法 LSRO 和 dMpRL 高 1.48% 和 5.9%。通过增加再排序技术，本章的模型在 rank-1 准确率达到 80.37%，mAP 也达到 86.38%，与 SVDNet[147] 相比性能不相上下（81.80% 和 84.80%）。如果不使用再排序技术，本方法在所有三个数据集中的表现都不如其他最先进的方法，这其中的一部分原因是在于实验中的基础网络的选择是一个相对简单网络，另外一部分原因是本章工作的重点是解决训练数据增强的伪标签问题而不是致力于明确处理诸如遮挡，尺度和错位等难以处理的情况。

3.4.3 消融实验

本节在 Market-1501 数据集上进行了消融实验，以评估本章提出的方法的每个组成部分对最终结果所带来的影响。可以注意到的是，当不使用 GAN 生成图像的情况下，本章的方法就退化成一个全监督的方法，而当无标签的数据被加入网络进行训练后，本方法就是变成一个半监督的方法。

表 3.10 消融实验结果

方法	rank-1	mAP	监督模式
Baseline	72.74	50.99	全监督
Center	79.45	57.25	全监督
FAPL-o	82.04	61.26	半监督
FAPL-d	**83.43**	**63.23**	半监督

中心损失 中心损失可以减少数据点之间的类内变化，在本章提出的伪标签方法中起着至关重要的作用，它可以保证学习的特征表示更加具有鉴别性[181]。在该实验中，仅仅使用有标签的数据用于训练以显示中心正则化的有效性。当加入了中心损失时

(表3.10中的第二行)，基准网络的 renk-1 准确率得到了 6.71% 的提升（从 72.74% 增加到 79.45%)，mAP 从 50.99% 增加 6.26% 达到 57.25%。这证实了中心损失对学习更具辨别力的表征的有着积极的影响。

伪标签 除了中心损失的有效性之外，在此基础上加入了本章所提出的两种不同编码的伪标签，分别表示成表3.10中的 FAPL-o 和 FAPL-d。可以观察到当由 GAN 生成的无标签数据通过生成伪标签来进行网络训练时，该网络在两个指标上都获得了进一步的性能提升，分别达到了 82.04%（独热编码）和 83.43%（分布式编码）的 rank-1 准确率，60.26%（独热编码）和 63.23%（分布式编码）的 mAP 值。

3.5 本章小结

本章在一种半监督的行人再识别设定下，提出了一种引入由 GAN 生成的图像进行训练的模型正则化方法，该方法主要通过为 GAN 图像生成伪标签来辅助训练。相较于现有的伪标签生成方法，本章提出的伪标签生成方法着重在于考虑无标签数据与有标签数据之间的特征相似性，并利用它们之间的相似性进行伪标签的生成。除此之外，本章还给出了两种标签的编码模式，分别是独热码和分布码，该两种编码模式在基于特征相似性的考量下可以在一个统一的学习框架中进行实现。实验结果表明，本章提出的方法在行人再识别任务方面的表现优于其他伪标签生成方法，并且与最先进的解决方案相比，达到了具有较强竞争力的行人再识别准确性性能。

4 基于分散度的无监督行人再识别

大多数行人再识别方法都是基于有监督的学习，例如前两章中所叙述的分别是全监督和半监督行人再识别任务。然而在行人再识别领域，带监督的学习就意味着需要对训练数据进行手动标注。然而，获取行人的身份标注不仅是很消耗人力物力资源的，并且标记大规模现实世界级的数据也是不切实际的。繁琐的数据标签获取用于行人再识别任务来进行学习使得其在现实世界场景中的部署变得更加困难。因此，如何在没有明确的监督下也能够让模型进行鲁棒的特征学习就成为了一个非常重要的课题。基于这样的情况，本章从聚类的角度出发，提出了一种简单但有效的无监督的行人再识别方法。具体来讲，本章借用统计学中的分散度的基本概念来实现一个稳健鲁棒的聚类标准。当在评估一个簇的分散度的时候，值得考虑的是应该是目标簇的紧凑性。如果一个簇有更小的分散度，那么意味着其越紧凑。当在评估簇与簇之间的分散度的时候，值得考虑的应该是簇与簇之间的分离程度。如果两个簇之间有更大的分散度，那么意味着其越分离。基于这种观察，本章提出了一种基于分散度的聚类（Dispersion based Clustering, DBC）方法，其可以更好地挖掘数据低层蕴藏的分布模式。本章提出的方法还可以自动确定独立数据点的优先级，并防止不良的群集的形成和蔓延。

4.1 引言

根据类别未知 (没有被标记) 的训练样本解决模式识别中的各种问题，称之为无监督学习。近些年，无监督学习与有监督学习一样，一直吸引着研究人员的注意力。行人再识别领域近年来年来已经取得了令人印象深刻的成果[27,80,147,183,185,190,191]。大多数现有的行人再识别模型都是在监督模式下使用带注释的 ID 标签进行训练[21]。因此，它们在实际应用中的部署通常受到缺乏大规模注释训练集的阻碍。然而，在复杂场景中获取手动身份标签是一项艰巨的工作，因此很多的工作对无监督的解决方案进行了研究。传统的无监督解决方案主要是基于手工特征[12,32,192]、显著性分析[38,193] 和字典学习[194]。与带监督的模型相比，这些无监督学习的初步尝试都往往得到比较低的性能。然而，在不同的照明和相机视角条件下，为不同相机拍摄的图像设计一个合适而通用的手工特征是一项具有挑战性的任务。如果没有成对标识标签，这些方法也无法明确地利用跨视图的判别信息。Farenzena 等人[32] 利用人物图像中的对称属性来处理视图差异。为了处理光照变化和杂乱的背景，Ma 等人[195] 提出将 Gabor 滤波器和协方差描述符结合起来。文献[196]中对 Fisher 矢量进行了探索，通过对于局部特征的高阶统计量来进行编码。最近，Xiao 等人[148] 也提出了半监督的行人检索 OIM 损失，它也可以用于无监督的行人再识别。然而，与监督学习方法相比，这些方法的性能远比有监督方法要弱得多。

在没有针对特定任务的标记数据的情况下，域适应（Domain Adaptation）通常是

一个具有潜力的方法，因为具有类似性质但来自不同域的标记数据是可用的。域适应的主要做法是对齐源域和目标域之间的特征分布[197–201]。Tzeng 等人[202] 引入了适应层和额外的域混淆损失，以学习具有语义意义和域不变性的表示。Long 等人[198] 则建议在深层网络中将所有任务特定层的隐藏嵌入到 RKHS 中，并匹配此空间中不同域分布的平均嵌入。在文献[203]中，Zhong 等人引入了异质同源学习，以对齐域分布。Ganin 等人[204] 利用对抗性学习来匹配源域和目标域之间的特征分布。最近，一些跨域转移学习方法[23,205] 已经对行人再识别领域进行了研究，以处理不同数据集之间的身份之间的错位。为了更好地弥合这一数据分布的差距，Wang 等人[206] 首先在源域上进行属性的训练，并学习身份和属性的联合特征表示。此外，一些工作使用生成对抗网络（GAN）生成增强图像以减少数据集差异[9,23]。Deng 等人[9] 则探索了图像自相似性和跨域不相似性以实现目标域图像的转换，Zhong 等人[23] 利用相机到相机对齐来执行图像转换。那些域适应方法都集中在目标域的标签估计上。

聚类作为一种重要的数据分析工具，也已在无监督的行人再识别中得到了应用。Fan 等人[24] 将域转移和聚类结合起来，应用于无监督的行人再识别任务。他们首先在外部带标记的行人再识别数据集上训练模型，可用该数据集得到一个良好的初始模型。之后，根据样本到各类簇质心特征之间的距离远近，逐步选择无标签的数据用于训练。但是，这项工作依赖于对身份总数的强烈假设。除了需要辅助数据集或假设的这些方法之外，Lin 等人[207] 还提出了应用自下而上的聚类框架，该框架根据某些标准对聚类进行分层组合并取得了比较好的结果。Lin 等人[207] 采用的合并标准是非常简单的最小距离标准。文献[207]中的自下而上的聚类方法本质上是一种层次聚类算法。这类算法的两个主要类别是凝聚聚类和分裂聚类。层次结构固有的性质表明用于簇合并或划分的标准是至关重要的，而这些标准通常被定义为（不）相似性度量。在文献[207]中，两个簇中图像之间的最小距离被当簇间相似性，作为衡量合并的标准。然而，该标准可能是有问题的，因为它仅考虑来自两个簇的一对图像，丢弃其他有用的线索。这种最基本简单的标准可能导致细长的簇，形成不正确簇从而导致性能变差。

在这里，本章试图通过探索特征空间中的数据之间的分散度来解决这个重要问题。一个好的聚类应该遵循以下两个基本属性，即簇内紧凑性和簇间良好分离程度。在统计学中，分散度指的是数据分布被拉伸或挤压的程度。其基本概念与聚类衡量标准相吻合，因此该分散度指标也能用来衡量一个聚类的质量。低簇内分散度和高簇间分散度是一个有效簇的标志，反之亦然。基于此，本章提出采用这种简单而优雅的标准作为凝聚聚类的合并规则来改善以往工作对于聚类标准的选择的缺陷。这项工作的主要贡献总结如下：

(1) 本章提出了在行人再识别任务中，使用数据点之间的分散度作为凝聚聚类的簇候选合并标准。该标准由两方面的考虑所组成，分别是簇间分散度和簇内分散度。簇间分散度主要衡量的是簇与簇之间的分离程度，而簇内分散度主要衡量的是候选簇本身的聚类质量。

(2) 该基于分散度的簇候选和合并标准主要有以下三个优点，分别是孤立数据点的自动优先排序，防止不良聚类的形成与蔓延和与网络学习有相互促进的作用。

(3) 不管在基于图片还是基于视频的行人再识别数据集上的实验结果表明，基于分散度的簇候选和合并准则都取得了比其他方法更高的性能结果。

4.2 聚类方法

4.2.1 聚类定义

将物理或抽象对象的集合分成由类似的对象组成的多个类的过程被称为聚类。由聚类所生成的簇（clusters）是一组数据对象的集合，这些对象与同一个簇中的对象彼此相似，与其他簇中的对象相异。对于一个高维空间中的数据点，其可以形成的簇可以描述为“该空间的连续区域包含相对高密度的点，通过相对低密度的点的区域与其他高密度区域分开。”接下来，本小节将尝试着用一些数学的公式来定义一下什么是聚类：

给定一个数据集 $\mathcal{X}$, 即：

$$\mathcal{X} = \{x_1, x_2, ..., x_N\} \tag{4.1}$$

定义 $\mathcal{X}$ 的一个 m 聚类 $\mathfrak{R}$, 将 $\mathcal{X}$ 分割为 m 个集合 $\mathcal{C}_1, \cdots, \mathcal{C}_m$，使其满足一下的三个条件：

- $\mathcal{C} \neq \emptyset, \quad i = 1, \cdot, m$
- $\cup_{i=1}^{m} \mathcal{C}_i = \mathcal{X}$
- $\mathcal{C}_i \cap \mathcal{C}_j = \emptyset, \quad i \neq j, \quad i, j = 1, \cdots, m$

另外，在簇 $\mathcal{C}_i$ 中包含的数据点彼此之间应该有更“相似”，与其他簇中的数据点则“不相似”。这些“相似”与“不相似”的定义则决定了聚类的好坏。也就是说，不同的相似度定义标准对于同一组数据来说，可能会导致完全不同的聚类结果。

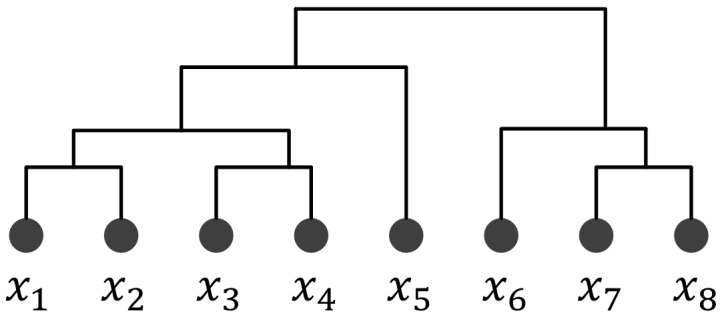


图 4.1 层次聚类树状图

4.2.2 聚类分类

聚类算法可以被视为通过仅考虑包含 $\mathcal{X}$ 中所有可能划分的集合的一小部分，就可以得到一个合理的聚类方案的方法。而正如上段所说，聚类的结果取决于特定算法和所使用的标准。因此，聚类算法也可以看成是一种学习过程，其目标是在没有任何外界信息辅助的情况下试图识别数据集下面的聚类的特定特征。聚类算法可以分为以下四个主要类别：

- **顺序算法**。这些算法目标是产生单个聚类。它们都属于一种非常简单快捷的方法。其中的大多数方法，所有特征向量按照一定的顺序输入给算法，可能是一次或几次。而最终结果通常取决于将向量输入给算法的顺序有关。根据所使用的距离度量，这些方案倾向于产生紧凑和超弹性或超弹性形状的簇。
- **层次算法**。这些方案进一步分为：*凝聚算法*。 这些算法在过程中的每一步产生一系列的簇数来逐步减少的待聚类的簇数量。具体地说，在算法的每一步，通过将前一步中得到的簇通过某些度量进行候选簇的筛选，并将候选簇进行聚类合并为一个簇。凝聚算法的主要代表是单一和完整的链接算法。*分裂算法*。这些算法与凝聚算法采用一种相反的方式，他们会在聚类的过程中逐步增加簇的数量。每一步中增加的簇来自于该步骤中簇的分裂。
- **基于损失函数优化的算法**。该类方法指的是通过使用损失函数 $\mathcal{J}$ 来定义聚类“合理程度”的方法。一般来说，簇的数量 m 保持固定。这些算法中的大多数都使用微分计算并在尝试优化 $\mathcal{J}$ 时产生的聚类。它们的终止条件是当 $\mathcal{J}$ 取得局部最优值的时候。该类别的算法也称为迭代函数优化方案。此类别包括以下子类别：硬聚类、概率聚类、模糊聚类、可能聚类、边缘检测算法等。
- **其他方法**。本类方法主要包括一些特殊的聚类技术。它们包括：分支和约束聚类算法、遗传聚类算法、随机松弛算法、谷点搜索算法、竞争学习算法和基于密度的算法等等。

4.2.3 凝聚聚类

本章提出的方法属于凝聚聚类算法的类别。下面本节将详细介绍一下凝聚聚类算法。凝聚聚类算法并不产生单一的簇，而是生成具有层次结构的簇。这种算法通常在社会科学与生物学等领域得到广泛应用。

如果生成的聚类 $\mathfrak{R}_i$ 中的每一个簇都是 $\mathfrak{R}_{i+1}$ 中簇的子集，即 $\mathfrak{R}_{i+1}$ 中的有些特定的簇是由 $\mathfrak{R}_i$ 中的簇聚合而成。分层聚类算法产生嵌套聚类的层次结构。更具体地来讲，这些算法涉及 N 个步骤，与数据集中的数据量一样多。在每个聚类步骤 t 中，基于在前一步骤 $t-1$ 产生的聚类获得新的聚类。整个凝聚聚类的过程可以参见图4.1很明显，该凝聚聚类算法有一个缺点，那就是它不能从一个“坏”的聚类中恢复，因为“坏”聚类

可能产生在更早的聚类层次上。这样，聚类过程中需要总共的聚类对的数量可以写成：

$$\sum_{t=0}^{N-1}\begin{pmatrix} N-t \\ 2 \end{pmatrix} = \sum_{k=1}^{N}\begin{pmatrix} k \\ 2 \end{pmatrix} = \frac{(N-1)N(N+1)}{6} \tag{4.2}$$

也就是说，凝聚聚类的运算复杂度在 $\mathcal{O}(N^3)$。另外，该运算复杂度还与相似性度量的定义相关。

4.3 基于分散度的无监督行人再识别方法

4.3.1 预备知识

给定一个无标签的数据集，$\mathcal{D} = \{x_i\}_{i=1}^{N}$，其中包含 N 张裁剪过的行人图片，此处的目标是在不借由任何的标签信息的情况下在数据集 $\mathcal{D}$ 学习一个特征嵌入函数 $\phi(x_i;\theta)$。其中的参数 θ 可以通过优化一个损失函数来完成。在训练完成之后，该特征提取器可以对查询集 $\{x_i^q\}_{i=1}^{N_q}$ 和候选集 $\{x_i^g\}_{i=1}^{N_g}$ 中的图片进行特征提取，并且利用这些特征进行基于距离的图像检索。两张图片之间的特征距离可以定义为 $dist(x_i^q, x_i^g) = \|\phi(x_i^q;\theta) - \phi(x_i^g;\theta)\|$。该距离大小可以体现为该图相对属于同一个人的可能性。即：具有越近的特征距离的图像对属于同一个的可能性就越高，具有越远的特征距离的图像对属于同一个人的可能性就越低。

全监督学习的情况下，每张输入图片 x_i 都会有其所属的行人 ID 标签 y_i。为了学习输入和输出之间的映射，通用方法是在特征嵌入函数后再加上一个分类器 $f(\phi;w)$，其中 w 是分类器的参数。因此，$\phi(x_i;\theta)$ 可以通过优化以下目标函数来实现：

$$\min_{\theta,w}\sum_{i=1}^{N} l(f(\phi(x_i^q;\theta);w), y_i) \tag{4.3}$$

其中 l 是分类的交叉熵（cross-entropy）损失。交叉熵损失的一个缺点是它并没有明确地最小化类内距离。为此，有人也提出了旨在实现类内紧凑性的中心损失（center loss）函数。

与中心损失函数类似，排斥损失（repelled loss）可以作为一个分类器，并且它具有通过基于特征相似性计算概率来共同考虑类间和类内差异的能力，其函数如下所示：

$$p(y|x,V) = \frac{\exp(V_y^T v/\tau)}{\Sigma_{j=1}^{N}\exp(V_j^T y/\tau)} \tag{4.4}$$

其中 τ 是一个温度参数，用于控制类别的概率分布的柔和度，v 是 $\phi(x;\theta)$ 经过 l_2 范式归一化过后的图像特征，而 V 是一个速查表（Lookup Table, LUT），该表中包含了每个类的质心特征。该速查表可以即时更新，可以避免特征提取的计算成本。

4.3.2 学习框架

将上述框架用于无监督行人再识别的主要瓶颈在于如何对未标记数据进行自动的标签分配。在这样的情况下，聚类是一种自然的选择，因为它旨在将同一组中的相似实

体组合在一起。在本章中，将提出一种新的基于分散度的层次聚类方法。两个簇之间的相似性/相异性度量的选择是本节提出的算法的关键。在行人再识别的任务中，重点在于识别相同身份的图像。该过程中应该考虑到簇的内部和外部相似性，这样才能得到一个比较合理的聚类。

给定在特征空间中散布的一个数据簇 C ，其分散度 $d(C)$ 可以被定义为簇内的平均成对距离，其定义以下：

$$d(C) = \frac{1}{n} \sum_{i,j \in C} dist(C_i, C_j) \tag{4.5}$$

其中 n 是簇 C 的元素数量。基于这样的定义，簇之间的分散度可以写成：

$$d(C_a, C_b) = \frac{1}{n_a n_b} \sum_{i \in C_a, j \in C_b} dist(C_{a_i}, C_{b_j}) \tag{4.6}$$

为了能够共同考虑簇内和簇间的分散，簇 C_a 和簇 C_b 之间的分散度可以表述为：

$$D_{ab} = d_{ab} + \lambda(d_a + d_b) \tag{4.7}$$

其中，为了符号使用的简单性，使用 d_{ab} 和 d_a 代替表示 $d(C_a, C_b)$ 和 $d(C_a)$，λ 是两个组件之间的权衡参数。公式(4.7)的前一项 d_{ab}，也就是簇之间的分散度，可以看成是簇与簇之间的不相似性度量。如果簇与簇之间的相似度很低的簇应该就可以考虑融合在一起，因为同一个行人的图像在特征空间中相互之间就应该有较低的不相似度。公式后一项 $d_a + d_b$，是两个候选簇的分散度之和，其可以充当正则化项的作用。该项的添加还有另外两个很重要的作用。第一个作用，它可以帮助确定独立数据点的优先级，以便在开始阶段进行合并；第二个作用在于它可以防止“坏”簇的形成。事实上，基于该公式的候选簇选择策略就是一方面需要考虑簇与簇之间的距离 ($\lambda = 0$)，并且另一方面也同时考虑单独簇自身的属性 ($\lambda \to +\infty$)。

4.3.3 矩阵更新

该聚类过程的输入是不相似性矩阵 $P(C)$，也通常被称为邻近矩阵。它是一个 $C \times C$ 的矩阵，其中 (i, j) 位置的元素即为 C_i 和 C_j 之间的簇间离散度 $d(C_i, C_j)$。$P(C)$ 可以通过首先计算图像成对距离矩阵来快速得到，该距离矩阵可以通过计算从深度网络获得的图像特征之间的外积来得到。在每一步的聚类过程中中，当选择合并两个簇时，不相似矩阵 $P(C)$ 的大小将会改变，变为 $(N-1) \times (N-1)$。在每一次的操作中，将会从不相似矩阵中删除对应的合并簇 C_a 和 C_b 的两行和两列列，并添加了一个新的行和一个新的列，其中的值都是新形成的簇 C_q 和旧簇 C_s 之间的更新过后的相似度。同样通过本章提出的分散度的定义可以计算出 C_q 和 C_s 之间的不相似性，表示如下：

$$d_{qs} = \frac{n_a}{n_a + n_b} d_{as} + \frac{n_b}{n_a + n_b} d_{bs}. \tag{4.8}$$

类似地，新形成的簇 C_q 的簇内分散度可以写成：

$$d_q = \frac{n_a d_a + n_b d_b + n_a n_b d_{ab}}{n_a + n_b + n_a n_b} \tag{4.9}$$

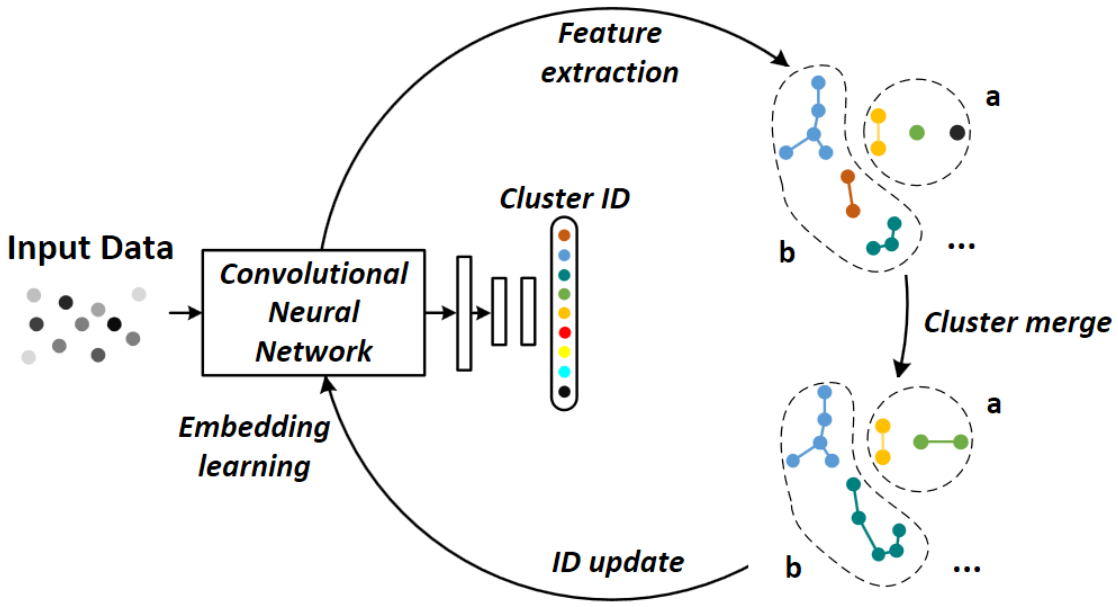


图 4.2 基于分散度的无监督学习总体框架

4.3.4 学习过程

基于分散度聚类学习方法的框架如图4.2所示。整个网络可以分为互相迭代的卷积神经网络训练学习和基于分散度的聚类两个步骤。在学习的最开始阶段，首先为训练数据集中的每个数据点 x_i 都分配一个唯一的标签 y_i。有了该初始标签，图像标签对 (x_i, y_i) 可以被输入到卷积神经网络进行分类学习，在训练的过程中可以使用反向传播的梯度更新其参数。在卷积神经网络训练完成之后，根据训练得到的卷积神经网络对所有图像进行特征提取，并在提取到的特征上利用公式(4.7)进行簇间不相似性计算。注意在此阶段，由于每个图片都单独属于一类，所以簇间不相似性跟样本的不相似性是一致的。接着根据所有的簇间不相似性按照从小到大的顺序进行排序并选出其中具有最小簇间相似度的 k 对簇进行聚类。k 是预定义的每轮聚类的合并数量。当两个簇被聚成一类之后，所有两个簇中的图像都将被分配一个相同的新的标签用于下一卷积神经网络的训练之用。该标签的分配规则可以用如下公式表达：

$$\mathcal{Y} = \{y_i = j \quad if \quad x_i \in \mathcal{C}_j\}_{i=1}^{N} \tag{4.10}$$

本章所提出的基于分散度的无监督学习的整个过程都在算法4中做了总结。

4.3.5 讨论

簇间分散度和簇内分散度的组合。 由公式(4.7)可见，该聚类标准由簇间分散度和簇内分散度组合而成。这样的设计对于聚类的过程来说，主要有两个好处：

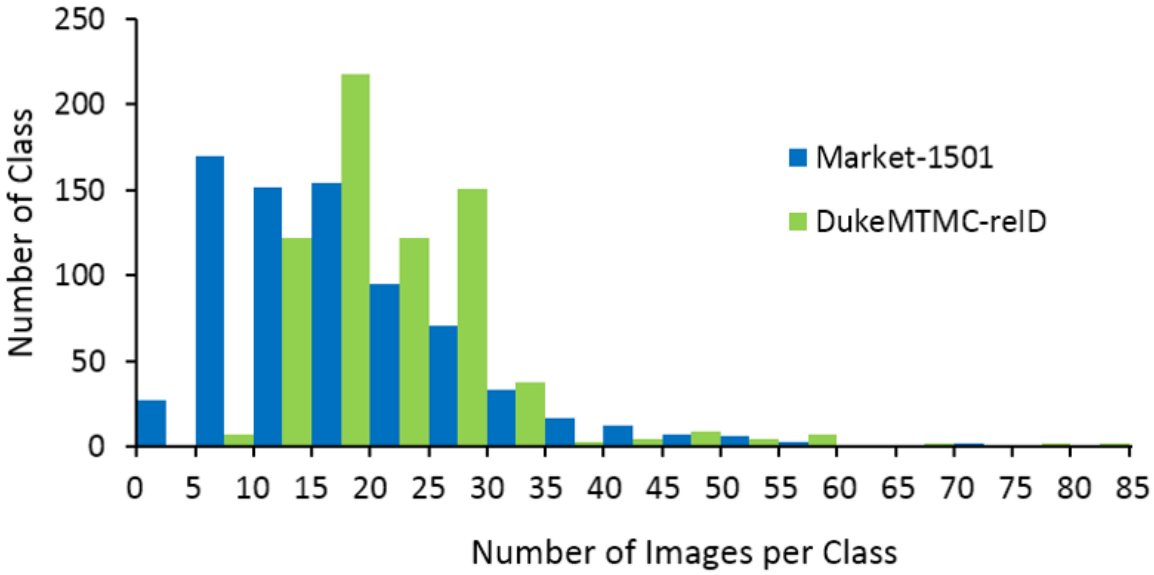


图 4.3 两个行人再识别数据库上的样本数量分布情况

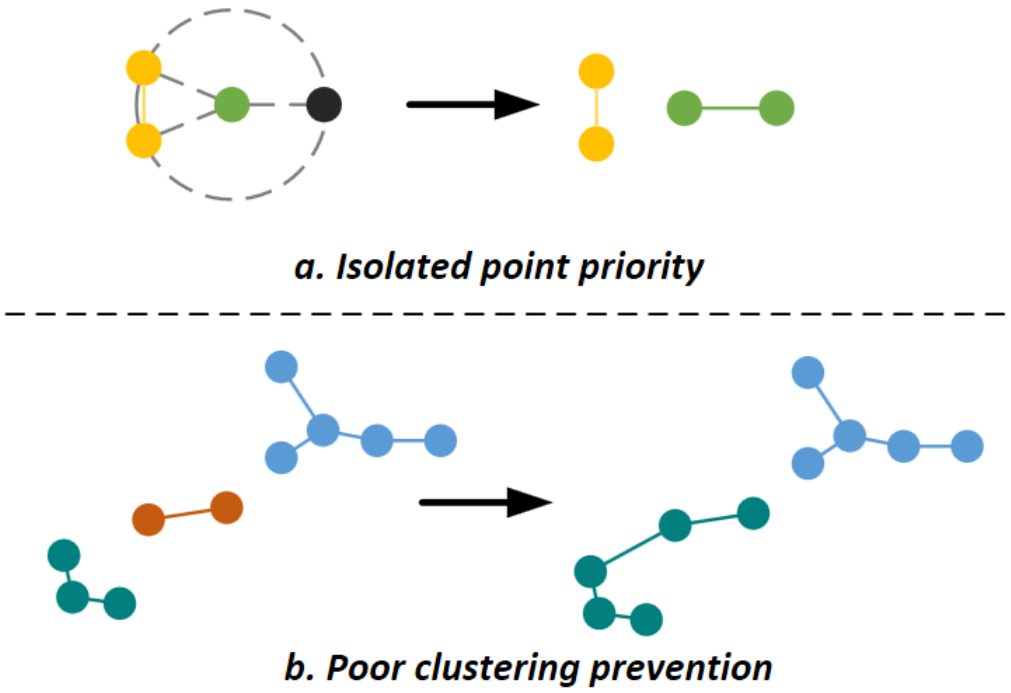


图 4.4 本章提出的方法的两个优点示意图

算法 4 基于分散度的误监督聚类特征学习方法

输入： 无标签训练数据 $\mathcal{D}=\{x_i\}_{i=1}^{N}$，聚类批系数 $m\in(0,1)$，权重参数 λ，卷积神经网络参数 $\phi(\cdot;\theta_0)$

输出： 优化后的模型参数 $\phi(\cdot;\hat{\theta})$

初始化: 初始训练标签 $Y=\{y_i=i\}_{i=1^N}$，簇数目 $C=N$，簇合并批数量 $k=m\times N$

while $C>k$ **do**

利用样本 $\{x_i\}$ 和其簇标签 $\{y_i\}$ 来根据公式(4.3)训练神经网络模型

计算簇间的不相似度矩阵 $\mathcal{P}(\mathcal{C})$

for 1:k **do**

根据定义的簇选择规则，公式(4.7)进行簇选择并合并它们

利用公式(4.8)和公式(4.9)进行簇间不相似性矩阵 $\mathcal{P}(\mathcal{C})$ 更新

$C\leftarrow C-1$

end for

利用新的簇信息 $\mathcal{C}$ 来更新训练样本类别标签 $\mathcal{Y}$

在验证集上验证网络模型的性能 $Perf$

if $Perf>Perf^*$ **then**

$Perf^*=Perf$

优化后的模型 $\phi(\cdot;\hat{\theta})$

end if

end while

返回: $\phi(\cdot;\hat{\theta})$

1）确立孤立点的优先级。 对于一个行人再识别数据集，一个合理的假设是它所包含的行人至少具有成对的样本数据。也就是说，在行人再识别数据集中，几乎不存在单独存在而不依附于任何其他点的数据，这是由于行人再识别任务的特性决定的。因此，独立的点在进行聚类的初始阶段应该具有更高的优先级，因为如果不优先考虑这些孤立点，他们很有可能在下一步的卷积神经网络的训练中被进一步推离与其属于同一个行人的其他数据点，因为他们具有不同的标签，而卷积神经网络就是被训练来将它们分开。当两个候选合并簇对具有相同的簇间分散度 d_{ab} 时，孤立点的优先级将被提升，因为它们具有较少（无）簇内分散度。一个孤立点优先级确立的示意图可以参看图4.4(a)。

2）防止坏聚类的形成和蔓延。 凝聚聚类的嵌套属性的一个缺点是无法从层次结构的先前级别中发生的“差”聚类中恢复[208]。当一个簇具有较高的簇内分散度的时候，它通常被认为是一个有效性较低的簇。使用簇内分散度 d_a+d_b 的好处就在于此。当先前步骤中的聚类发生了比较差的簇的时候，在本轮的聚合候选簇的选择的过程中，那些具有高簇内分散度的将由于其的高簇内分散度而被降低合并优先级，尽管他们可能具有更小的簇间距离。一个阻止坏聚类形成和蔓延的示意图可以参看图4.4(b)。

与相似工作的比较 本章的工作与 BUC[207] 具有一定的相似性，都是采用的聚合聚类的框架来完成无监督行人再识别的任务。然而，该方法和本方法在簇合并标准的方面存在着很大的差异。Lin 等人[207] 采用跨群集样本之间的最小距离来衡量它们的不相似性。众所周知，这种单链接算法具有链接效应，即簇间不相似性 d_{qs} 取自从 d_{as} 和 d_{bs} 中的较小者，即 $d_{qs} = min\{d_{as}, d_{bs}\}$。这也就意味着它有着倾向于形成细长的簇的趋势。这种被拉伸的簇的分布情况会对下一阶段的神经网络模型的训练产生负面的影响，因为卷积神经网络的训练就是形成更加紧凑的簇，而这种细长的簇则使得神经网络模型的训练需要更多的迭代次数，两者互相是抵触的。基于行人再识数据库的训练样本均匀分布在行人 ID 之间的假设，Lin 等人[207] 提出使用簇中数据的数量作为一种多样性正则化项，其考量是在于不同的行人 ID 之间的图像样本数量差异性应该不大，然而这种假设并不一定能够满足。图4.3中展示了在 Market-1501 数据集和 DukeMTMC-reID 数据集中的各类别样本数量的比较，可以看到的是，在行人再识别数据集上，样本数量的呈不均匀分布状，并通常带有一个很长的尾巴，因此数据均匀分这样的假设并不能成立。与文献[207]不同的是，本章提出的标准通过对个体数据点对之间的距离进行衡量，可以更好地利用这种数据之间的关系；另外一点在于，本章认为图像样本的个数并不应该是一个值得考量的东西，相反来说，应该更加注重于样本之间的相似度，如果样本之间具有足够的相似度，尽管其包含的数可能比较多，还是应该将其认为是属于同一个行人身份的。综合而言，本章提出的簇合并标准可以帮助更好地形成紧凑和良好分离的聚类。

4.4 实验与结果分析

4.4.1 数据集

在本章的无监督行人再识别的训练学习中，采用了四个标准的行人再识别数据集，其中包含两个基于图片的数据集，即 Market-1501[8] 和 DukeMTMC-reID[116]；还有两个基于视频的数据集，分别是 MARS[14] 和 DukeMTMC-VideoReID[25]，整体的数据库训练集测试集的情况可参照表4.1。基于图片的行人再识别数据集 Market-1501 和 DukeMTMC-reID 已经在之前的章节1.3中作了详细的介绍，下面将介绍一下两个基于视频的行人再识别数据库，分别是 MARS 和 DukeMTMC-VideoReID。

MARS 行人再识别数据库[14] 该数据集是 Market-1501[8] 的一个扩展版本。在行人视频的采集过程中，共采用了 6 个摄像机，其中包含 5 个分辨率为 1080×1920 的相机还有一个 640×480 的相机。总共采集了 1261 个行人的 20715 段视频。其中，625 个行人被用作训练集，636 个行人用作测试集。

DukeMTMC-VideoReID 行人再识别数据库[25] 该数据集也是从 DukeMTMC-reID[116] 的扩展。该训练集包含了 702 个训练行人，每个行人只有一段视频片段。测试集同样也包含了 702 个行人。除此之外，该测试过程中还加入了 408 个干扰行人。总的来说，训练集有 2196 个视频片段，测试集总共包含了 2636 个视频片段。

表 4.1 基于图片和视频的行人再识别数据库的对比

数据库	类别	训练集		测试集			
		ID 个数	样本个数	查询集		检索集	
				ID 个数	样本个数	ID 个数	样本个数
Market-1501[8]	图片	751	16,522	750	3,368	750	19,732
DukeMTMC-reID[116]	图片	702	16,522	702	2,228	1,110	17,661
MARS[14]	视频	625	8,298	626	1,980	636	12,180
DukeMTMC-VideoReID[25]	视频	702	2,196	702	702	801	2,636

(a) MARS (b) DukeMTMC-VideoReID

图 4.5 基于视频的行人再识别数据集样例

4.4.2 实验设置

训练设置 为了使用上述的行人再识别数据集进行无监督学习，需要对如上的一些数据集做一些训练设置的修改，具体的训练设置修改如下。对于基于图像的行人再识别数据集，即 Market-1501 和 DukeMTMC-reID，在实验中移除了所有图片的身份 ID 标签，并为每一张图片初始化一个不同的标签，即：标签数等于训练集图片数，其他的训练设置保持不变。类似地，对于基于视频的行人再识别数据集，即 MARS 和 DukeMTMC-VideoReID，由于该训练集的最小单位是一个行人的一段视频，于是将每个视频片段作为一个不同的行人来进行标注，即：标签数等于训练集的视频数。在此需要再次说明的是，本章并没有使用任何的额外的注释信息来进行模型的初始化或辅助之后的无监督学习过程。

评估设置 在网络训练和聚类学习完成之后，得到的卷积神经网络模型将用作特征

提取器。对于基于图片的行人再识别数据库的图片输入，其提取的行人特征表示就是从卷积神经网络的倒数第二层的输出，而对于基于视频的行人再识别数据库的视频输入，其提取的行人表示则为其所有帧的特征的平均值。在得到所有的行人表示之后，这些描述符将用于基于欧几里德距离的检索。在本章中，将仍使用 rank-k 准确率和平均精度（mAP）来评估所提出的方法。Rank-k 反映的是检索的精度，而 mAP 则反映了整体精度和召回率。

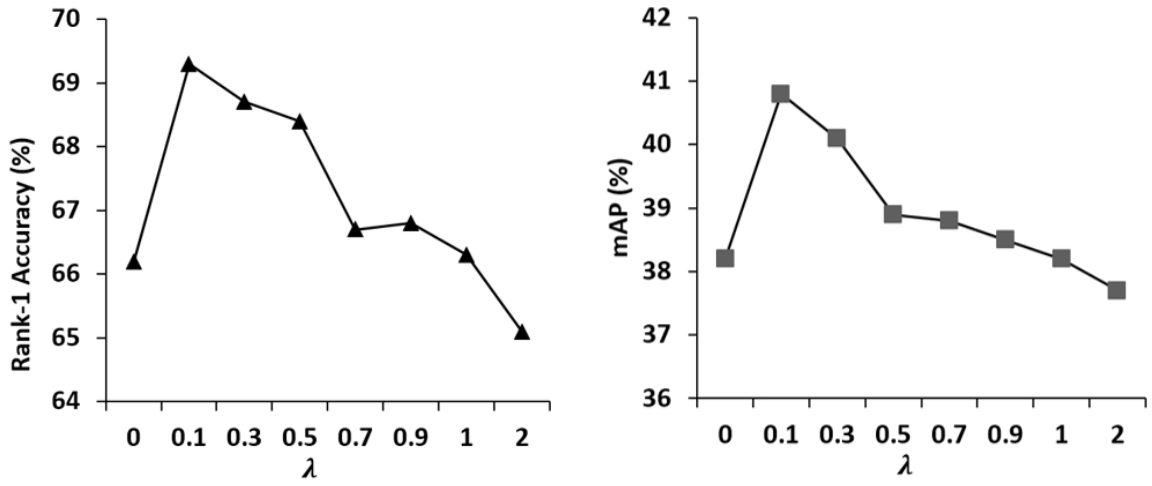
4.4.3 实现细节

在本章的实验中，对于卷积神经网络部，广泛使用的 ResNet-50[124] 被选做本方法的骨干架构，并使用了其在 ImageNet[209] 上预先训练的参数来进行模型初始化。除此之外，在 ResNet-50 倒数第二层的基础之上添加了两层的全连接层，用于更小的特征学习。最后的分类层是根据公式(4.4)进行实现的，其中温度函数 τ 设置为 0.1。在后续的实验过程中，除了明确说明之外，在所有的数据集（包括基于图片的和基于视频的）上使用的是完全相同的一组超参数。对于卷积神经网络模型的训练，总训练周期被设定为 20，批大小为 16，dropout 为 0.5，m 为 0.05。该卷积神经网络模型的参数优化方法选择的是随机梯度下降（SGD）优化，动量设定为 0.9。参数的学习率初始化为 0.1，并在 15 个训练周期后降低到 0.01。在聚类过程中，公式(4.7)中的权衡参数 λ 设置为 0.1。

4.4.4 算法分析

时间复杂度 在计算机科学领域，算法的时间复杂度是一个函数，它定性描述了该算法的运行时间。算法的时间复杂度是衡量一个算法效率的重要保证。在本节，将分析本章所提出的算法的时间复杂度。由于本章提出的算法可以看成是一个凝聚聚类算法的候选簇选择标准，于是着重分析这部分所带来的的时间复杂度。首先，该算法需要计算出整个数据集中任意两元素之间的距离，可以写成 $\mathcal{O}(N^2)$，其中 N 为数据元素的个数。其次，需要计算的是任意两个簇之间的分散度和簇内分散度，它们所需要的时间复杂度为 $\mathcal{O}(C^2)$，其中 C 代表的是当前状态下的簇的数量。在计算得到簇间和簇内分散度之后，需要根据公式(4.7)按照从小到大的顺序进行排序，需要 $\mathcal{O}(C\log C)$。最后在每次聚类过程需要进行 k 次的簇融合，所以需要 $\mathcal{O}(kC)$ 的时间复杂度。因此，总的来说，单个聚类总过程所需要花费的时间复杂度为 $\mathcal{O}(N^2+C^2+C\log C+kC)$。可以看出的是，在一段时间的迭代之后，$C$ 和 k 的值都是小于 C 的，并且并不需要将所有的簇都最终聚到一个类上，所以，整个时间复杂度可以看成 $\mathcal{O}(N^2)$。实际上，所有不相似性距离都是可以通过 GPU 上的矩阵操作来计算的，能够比较有效率地执行。

平衡参数 本节对公式(4.7)中的正则化参数 λ 进行了分析研究。λ 是用来平衡簇间分散度与簇内分散度的权重。在 Market-1501 数据集上对 λ 进行的实验结果在图4.6中进行了展示。可以看出，随着 λ 的值从 0 开始慢慢增加的时候，rank-1 准确率和 mAP 值都首先有了一定程度的上升，在 $\lambda=0.1$ 的时候，rank-1 准确率和 mAP 值同时达到了峰值，接着，随着 λ 的继续增加，两种性能指标都开始慢慢下降。当 $\lambda>1$ 的时候，



(a) Rank-1 (b) mAP

图 4.6 Market-1501 数据集 λ 参数实验结果

rank-1 准确率和 mAP 值都比当 $\lambda = 0$ 是还要低。这整个过程相对来说是比较合理的，因为该参数的大小可以被解释为簇筛选的偏好问题。也就是说，当 λ 很小时或接近于 0 的时候，整个选择的标准是更加倾向于直接选择在特征空间中簇间距离比较小的候选簇，即：空间上接近的聚类；而当 λ 开始增加的时候，整个选择的标准就相对地把候选簇本身的的簇内分散度考虑进来。而当 λ 过大的时候，整个选择标准完全选择簇内分散度更加紧凑的簇而不再考虑它们之间的分散度，这对于聚类来说是不合理的，也就导致了最终聚类性能的下降。

性能稳定性分析 本节进一步用对比实验分析了本章提出的方法的性能稳定性。在性能稳定性分析部分，主要分析算法的以下两个方面，分别是收敛速度和鲁棒性。好的算法需要有更好的收敛速度以及更强的鲁棒性。整个对比实验在 DukeMTMC-VideoReID 数据集上进行，对比方法为 BUC[207]，在实验过程中，所有的参数设置都一样，不同的仅仅是簇选择标准而已。图4.7为实验对比结果，图中给出的是在整个训练过程中，随着迭代次数的叠加，在测试集上的性能表现曲线。

收敛速度 收敛速度分析是算法分析的一个重要组成部分。一个好的算法能够更快地收敛表明其能够更快地寻找到数据中隐含的规律。在图4.7中，可以看到的是在迭代次数比较少，训练刚开始的时候（训练次数小于 10 的阶段），BUC[207]（蓝色方块）和本章所提方法（橙色三角）的性能曲线相互交叠，并没有拉开太大的差距，而之后（训练次数大于 10），可以看到本章算法有了一个比较明显的上升的趋势，并花了 2 个训练周期，最终在第 12 次的迭代训练中达到了性能的顶峰，相比较而言，BUC[207] 在第 10 次迭代之后也有了性能的提升，但并没有很明显，最终在第 13 次迭代中达到了它的性能顶峰。对两段性能表现的分析如下：在聚类迭代的前期的时候，两种方法都是首先选择的是小的簇在融合，更多的是形成一些基础的类别信息，类间的区分也比较简单。所以导致两种方法的性能都在差不多的阶段。而当聚类到一定程度的时候，聚类的选择需

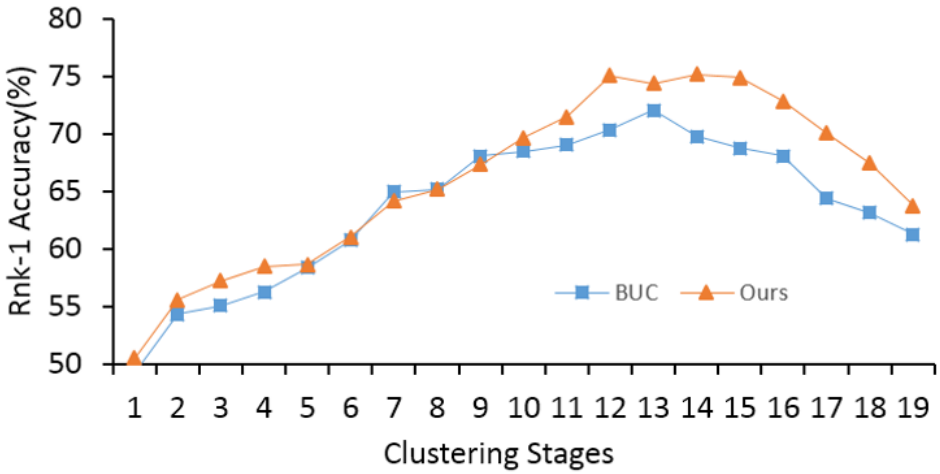


图 4.7 本章方法与相似方法的稳定性对比

要更加精细地选择，而此时本章的方法能够根据簇内分散度和簇间分散度的综合考虑来选择更加符合实际情况的候选类，所以本章的方法比 BUC[207] 提前了一个迭代周期完成。注意这里虽然只是提前了一个迭代，但每个迭代的周期都需要合并 $k = N \times m$ 个簇，在该数据集上就是要再合并约 110 个簇，如果是数据集 DukeMTMC-reID，则 1 个迭代周期就代表着需要额外进行约 830 次合并操作。

鲁棒性 算法性能的另外一个评价指标就是算法的鲁棒性。不同于算法在不同数据集上的性能表现的鲁棒性（该类鲁棒性可以在4.4章节看到），在这里所讨论的鲁棒性主要指的是对于聚类数目的鲁棒性。具体来说，指的是不同的目标聚类数目对于性能的影响的情况。在图4.7中，可以看到 BUC[207] 在第 13 次迭代的过程中取得了最高的性能结果，而紧接着的几次迭代，其性能就开始逐步呈下降趋势；反观本章提出的方法在第 12 次迭代就取得了最好的效果，然后在接下来的 3 次迭代直到第 16 次才有了较为明显的性能下降。这个现象充分说明了本章提出的方法在性能最高的时候对于聚类数目的不敏感性，具有更强的鲁棒性。除此之外，可以看到在之后的迭代步骤中，本章所提出的方法的性能都在 BUC[207] 的性能之上，从另一个方面说明了其鲁棒性。

表 4.2 聚类准则各组成部分的有效性实验

方法	Market-1501		DukeMTMC-reID		MARS		DukeMTMC-VideoReID	
	rank-1	mAP	rank-1	mAP	rank-1	mAP	rank-1	mAP
BUC^{-} [207]	62.9	33.8	41.3	22.5	55.5	31.9	60.7	50.8
BUC[207]	66.2	38.3	47.4	27.5	61.1	38.0	69.2	61.9
DBC^{-}	66.2	38.7	48.2	27.5	59.8	37.2	71.8	63.2
DBC	**69.2**	**41.3**	**51.5**	**30.0**	**64.3**	**43.8**	**75.2**	**66.1**

4.4.5 消融实验

为了更加直观地看到本章所提出的基于分散度的聚类标准的有效性，本节对其进行了消融实验。该消融实验的设定是逐步削减分散度标准的各个组成部分，以达到观察其各个部分对实验结果的影响。表4.2总结了在所有数据集上进行消融实验的数值结果并与 BUC[207] 作出了对比。

簇间分散度的有效性 在本节中，首先需要评估的是使用簇间分散距离作为簇间不相似性度量的有效性。在该实验中，去除了公式(4.7)中的第二项 $\lambda(d_a + d_b)$（即类内分散度），仅仅使用 ab 来作为最终的选择标准。为了比较的公平性，实验中也去除了 BUC[207] 中的簇内样本量的附加项。他们分别在表4.2中以 DBC$^-$ 和 BUC$^-$ 标识出来。可以看到的是，在所有四个数据集中，DBC$^-$ 的 rank-1 准确度比 BUC$^-$[207] 的都要高大概 6% 左右，同样地，mAP 值也要平均高出 7%。这样的性能增长足以表现出，本章所提出的簇间分散度的指标是一个很好的评判标准，原因在于 BUC[207] 仅仅考虑了单对样本之间的距离，而簇间分散度则是综合考虑了所有的样本对之间的关系。另外一个值得一提的是，本章方法 DBC$^-$ 在不需要正则项的情况下，就取得了与 BUC 全模型相近的结果，更加说明了该分散度对行人再识别任务的有效性。

簇内分散度的有效性 本节将进一步研究了簇内分散度正则化项的影响。完整模型 DBC 的性能表现在表4.2的最后一行列出，可以看到的是，正则化项的引入有助于提高算法的性能。在 Market-1501 上，rank-1 的准确度从 66.2% 增加到 69.2%，mAP 值从 38.7% 增加到 41.3%。并且在其他的数据集上在都表现出了性能的增加，其增加幅度都在 3% 到 5% 左右，这样的性能提升，说明了簇内分散度的正则项对于簇间分散度项有很好的互补作用。当在实验中把簇间分散度和簇内分散度组合在一起的时候，本章的模型可以取得最高的性能表现。

聚类结果的可视化 为了更加可视化地观察和研究通过本章提出的基于分散度的聚类方法在行人再识别数据库上得到的效果，本节在一个缩减数据集上做了实验并且利用 T-SNE[210] 对聚类结果进行了可视化。该缩减数据集是从 Market-1501 数据集中随机抽取了 100 个行人，其中包含 1657 张行人图片组合而成的。在该缩减数据集上，利用本章的方法对其中的数据进行了聚类。其聚类的可视化结果可参照图4.8。在图中，上半部分展示了经过降维后在二维空间中的聚类结果。不同的颜色代表不同的 ID 的行人样本，可以看到的是，大多数的行人样本都按照行人 ID 进行了比较好的聚类，其中一个比较好的聚类结果如右下深蓝色点所代表的行人，该类别包含的都是同一个 ID 的行人样，这样的聚类结果也就表明了本章提出的聚类方法的有效性。而另外一个比较不好的聚类结果在左下角进行了展示，其中两个不同 ID 的行人图片被聚到了同一个类中，经过 ID 样本的展示可以看到，这两个 ID 的行人样本之间存在着相当大的相似性，比如都是白色短 T 恤，黑色运动裤。这样的聚类结果也是情有可原的。也就是说当两个样本之间的差异性比较小的时候，是可能会导致不同的 ID 被聚到一起。这从另一方面也说明了，在相似性特别高的样本存在的情况下，能够获得一定的标签信息是很有必要的。

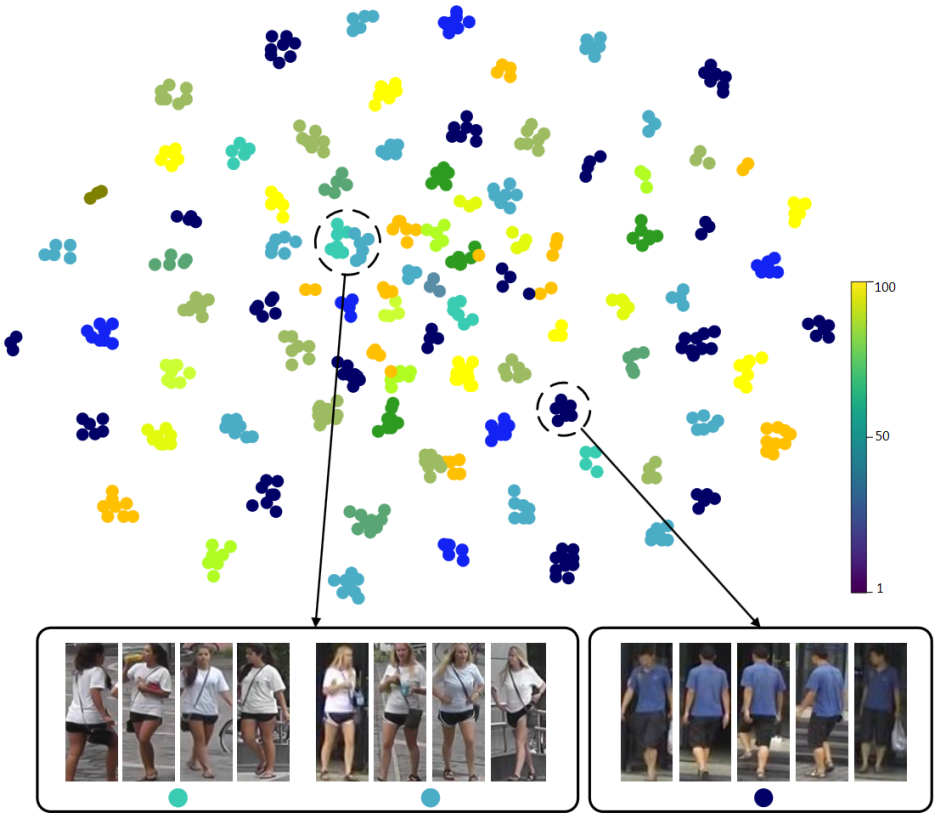


图 4.8 本章所提出的方法的聚类效果示意图

表 4.3 Market-1501 数据集实验结果

方法	标签	Market-1501			
		rank-1	rank-5	rank-10	mAP
BOW[8]	**None**	35.8	52.4	60.3	14.8
OIM[148]	**None**	38.0	58.0	66.3	14.0
UMDL[205]	Transfer	34.5	52.6	59.6	12.4
PUL[24]	Transfer	44.7	59.1	65.6	20.1
EUG[25]	OneEx	49.8	66.4	72.7	22.5
SPGAN[9]	Transfer	58.1	76.0	82.7	26.7
TJ-AIDL[206]	Transfer	58.2	-	-	26.5
BUC[207]	**None**	**66.2**	**79.6**	**84.5**	**38.3**
DBC	**None**	**69.2**	**83.0**	**87.8**	**41.3**

4.5 实验

接着，本节在基于图像和基于视频的行人再识别数据集上进行了实验，并将本章提出的方法与其他方法进行比较。所有的比较结果都在对应的表中展示出来。在这里需要说明的是各表中的“标签”列的值所代表的含义。在实验结果表的“标签”列一共出现了以下几种类别，分别是“None”代表实验过程中完全没有使用任何额外的标注信息；“Transfer”则表示的该方法或在开始或在过程中使用了其他的带标注的行人再识别数据库；“OneEx”则代表的是该方法是属于单样本学习的方法，即训练过程中对每个类别的行人有了单个的样本标注；“Camera”则表示的是，虽然没有使用任何的行人 ID 标签，但是使用了图像的摄像头属性，即哪些图片或视频段是从同一个摄像机下获取到的。在介绍完这些之后，将对本章所提出的方法与其他方法的得到的结果进行比较。

表 4.4 DukeMTMC-reID 数据集实验结果

方法	标签	DukeMTMC-reID			
		rank-1	rank-5	rank-10	mAP
BOW[8]	**None**	17.1	28.8	34.9	8.3
OIM[148]	**None**	24.5	38.8	46.0	11.3
UMDL[205]	Transfer	18.5	31.4	37.6	7.3
PUL[24]	Transfer	30.4	46.4	50.7	16.4
EUG[25]	OneEx	45.2	59.2	63.4	24.5
SPGAN[9]	Transfer	46.9	62.6	**68.5**	26.4
TJ-AIDL[206]	Transfer	44.3	-	-	23.0
BUC[207]	**None**	**47.4**	**62.6**	68.4	**27.5**
DBC	**None**	**51.5**	**64.6**	**70.1**	**30.0**

4.5.1 基于图像的行人再识别数据库

Market-1501 数据集上的实验 在 Market-1501 数据集上的实验结果可以参照表4.3，可以看到的是本章所提出的方法取得了最高的 Rank-1 准确率以及 mAP 值，它们分别是 69.2% 和 41.3%。与之最接近性能是 BUC[207]，其 rank-1 准确率和 mAP 值，分别是 66.2% 和 38.3%，都比本章提出的方法低 3%。其中完全使用无监督学习机制的还有 BOW[8] 和 OIM[148]，但其都属于一些比较简单的手工特征方法，所以取得的性能表现就会相对使用深度网络来说略显欠缺。

DukeMTMC-reID 数据集上的实验 在数据库 DukeMTMC-reID 数据集上的实验结果可以参照表4.4。同样的，本章所提出的方法在该数据集上也表现出了很高的性

能，该方法在 rank-1 准确率上首先超出了 50%，以及在 mAP 值上超出了 30%。可以看到性能结果接近的除了 BUC[207] 之外，还有 SPGAN[9] 和 TJ-AIDL[206]。SPGAN[9] 通过生成对抗网络（GAN），将未知标签的数据能够投射到有标注的数据集上去，从而能够实现标签的分配。而 TJ-AIDL[206] 则是利用了额外的属性（attribute）信息来训练一个能够同时包含身份信息和属性信息的特征表示从而进行了域迁移。虽然这些方法都使用了额外的数据以及标注，但它们的性能表现都远远低于本章方法，究其原因还是在于，对于行人再识别来说，进行域迁移的数据基础比较低，因为不同的数据集中行人的 ID 往往是没有重叠的。

表 4.5 MARS 数据集实验结果

方法	标签	MARS			
		rank-1	rank-5	rank-10	mAP
OIM[8]	**None**	33.7	48.1	54.8	13.5
DGM+IDE[211]	OneEx	36.8	54	-	16.8
Stepwise[212]	OneEx	41.2	55.5	-	19.6
RACE[213]	OneEx	43.2	57.1	62.1	24.5
DAL[214]	Camera	49.3	65.9	72.2	23.0
BUC[207]	**None**	61.1	**75.1**	**80.0**	38.0
EUG[25]	OneEx	**62.6**	74.9	-	**42.4**
DBC	**None**	**64.3**	**79.2**	**85.1**	**43.8**

表 4.6 DukeMTMC-VideoReID 数据集实验结果

方法	标签	DukeMTMC-VideoReID			
		rank-1	rank-5	rank-10	mAP
OIM[8]	**None**	51.1	70.5	76.2	43.8
DGM+IDE[211]	OneEx	42.3	57.9	69.3	33.6
Stepwise[212]	OneEx	56.2	70.3	79.2	46.7
BUC[207]	**None**	69.2	81.1	**85.8**	61.9
EUG[25]	OneEx	**72.7**	**84.1**	-	**63.2**
DBC	**None**	**75.2**	**87.0**	**90.2**	**66.1**

4.5.2 基于视频的行人再识别数据库

MARS 数据集上的实验 在数据库 MARS 数据集上的实验结果可以参照表4.5。在此数据集上，可以看到属于单样本学习的方法 EUG[25] 在 rank-1 和 mAP 值上都超过了属于完全无监督的方法 BUC[207]1.5% 和 4.4%，这说明了提供了一个样本标签之后的网络能够有一个比较好的初始值，比完全无监督的存在一些优势。然而，本章提出的方法在同样无监督的情况下却超出了 EUG[25]，其原因可能是在单样本情况下，无标签的样本在被根据特征相似度逐步选择之后，由于网络训练并不足够，所以对于特征相似的行人并没有得到很好的区分。而本章的方法并不要求不同的行人一定要不同的类别，而是从足够相似则属同一人的角度去放宽了约束，反而取得了更好的效果。在 MARS 上，Rank-1 准确率和 mAP 值分别是 64.3% 和 43.8%。

DukeMTMC-VideoReID 数据集上的实验 在数据库 DukeMTMC-VideoreID 数据集上的实验结果可以参照表4.6。在 DukeMTMC-VideoReID 数据集上，本章提出的方法体现出了更高的性能，其 rank-1 准确率和 mAP 值分别是 75.2% 和 66.1%。对比于基于单样本学习的 Stepwise[212] 算法来说，rank-1 准确率和 mAP 值的差距在 20% 左右，其算法还利用了互惠近邻来优化结果。

综合以上的多个数据集的实验结果证明本章提出的基于分散度的聚类算法对不同数据分布具有更稳定的泛化能力，从而该方法的稳定性和有效性得到了更好的证实。

4.6 总结

本章提出了基于分散度的聚类算法来处理无监督的行人再识别的任务。从聚类候选簇选择的角度来说，要形成一个良好的聚类，它一般都需要具备两个有利条件，单个簇的内部完备性，簇与簇之间良好的分离程度。本章所提出来的聚类算法通过簇内分散度来限制单个簇的完备性，通过簇间的分散度来保证良好的簇间分离程度，最终综合地考虑了这两个属性。除此之外，该基于分散度的候选准则还考虑到行人再识别数据库的数据分布特殊性，即属于同一类行人图片的大多能够较好地聚集在一起，该方法能够良好的对潜在的数据分布进行建模还表现在其具有的两个特性，分别是单独点的处理优先级以及防止坏聚类的形成。由于行人再识别一般不会出现单张图片的情况，所以对于单独点需要被赋予较高的优先级，从而保证在特征空间本来有可能有一些差距的同一行人图片能够优先被选择合并，因为卷积神经网络的训练会使原本远距离的样本更远；而一旦有了坏聚类的形成的时候，其往往会倾向于吞并其他的簇，所以簇内的分散度限制其无止境地合并其他簇导致聚类结果的继续恶化。另外，所提出的分散度模型与卷积神经网络的损失函数训练起到了相互促进的作用，都是在要求同一类别的样本在特征空间下能够更好地聚在一起，而不同类别的样本之间的距离能够拉开。最终，在实验数据集上进行了大量消融实验以及性能比较，经过对结果的分析，充分地证明了本章所提出的基于分散度的聚类方法在取得领先性能的情况下仍具有良好的稳定性和泛化能力，并且拥有

更快的收敛速度。

5 总结与展望

5.1 本文工作总结

本文基于深度卷积神经网络的特征学习表示能力，针对行人再识别特征学习中存在的问题展开研究和探讨，提出了在全监督情况下，针对网络特征学习的鲁棒性不够的问题，提出了一种互补特征的提取方法；在半监督情况下，针对如何利用生成对抗网络生成的无标签合成图片来对模型进行正则化操作，提出了一种基于特征相似度的伪标签生成算法；在无监督的情况下，针对如何利用无标签的行人数据，提出了综合考虑簇间和和簇内分散度的聚类方法。本文的主要研究内容、获得的结论以及创新之处总结如下：

(1) 基于互补特征的行人再识别方法从一个新的角度来解决深度学习网络所存在的特征过于全局化的问题，提出了一种能够根据现有全局特征自动学习出与之相互补的局部特征来提高特征的鲁棒性。所提出的互补特征提取网络类似于一个孪生结构，差别在于本网络两分支的输入是同一张图片，其中一分支根据其输入的标签信息进行全局特征的学习，该学习到的全局特征则经过一个特征掩码网络输入到第二个分支，在全局信息作为输入的基础上，此分支能够学习到互补的特征。为了更加保证两支网络的特征互补性，文章又提出了一个成对排序损失函数，通过它来引导和限制互补特征的学习。互补特征的组合大大提升了行人再识别的准确率。

(2) 基于模型正则化的行人再识别方法探究了如何在有标签数据量不足的情况下，使用生成对抗网络来生成合成图片作为数据增广的方式。文章首次提出了基于特征相似性的伪标签生成方法，该方法具有的优势在于更好地衡量合成样本与现有样本之间的标签关系，并且可以利用该方法在同一种框架下生成两种不同的伪标签编码模式。

(3) 基于分散度的行人再识别方法在面对行人再识别数据集打标签中存在的难度，解决如何利用无标签的行人数据来进行行人再识别。文章采用了一种凝聚聚类的方法来进行标签分配，并且提出了一种基于分散度的候选簇选择标准。该分散度包含了簇间的分散度以及簇内的分散度，簇间的分散度主要考虑两个簇之间的分离度，簇内分散度主要考虑的是簇内的紧凑性。相较于其他的选择标准来说，基于分散度的标准有多重优点，表现在其能够优先选择单独点以及阻止坏簇的形成。此外，该选择标准与神经网路训练损失函数之间能够起到互相促进的作用。通过实验证明了该选择标准相较于类似工作有更快的收敛速度以及更好的泛化能力和稳定性。

5.2 未来工作展望

特征提取是模式识别系统中的一个重要环节，对于高维数据如何提出更有效的特征提取准则一直是模式识别领域一个重要的研究方向。同样地，行人再识别领域也面临着相同的问题。近年来，带标签的数据集的规模的显著增加的趋势使得深度学习方法能够再行人再识别领域被广泛应用。深度学习是数据驱动的，没有大量的数据就很难学习出良好的特征提取模型。然而大规模的行人再识别数据的收集则更加具有挑战性，因为除了需要绘制行人的标定框之外，还必须人工为他分配一个身份 ID。而这个 ID 的分配在多人协作标注时需要很高的交流成本。本文基于上述的两个主要问题，分别对特征提取以及减少对标注数据的依赖两个方面进行了相关的探索研究并提出了一些方法。然而，鉴于作者的学识和时间有限，很多与此相关的工作还没有深入做下去。基于本文以及作者对本学科一些相关问题的理解和认识，提出一些笔者认为尚未解决但是亟待解决且非常有意义的问题。

(1) 对于鲁棒特征提取的问题，除了本文所提出的互补鉴别特征的挖取之外，还可以使用的是困难样本学习。困难样本学习的宗旨是把分类器错误分类的样本 (hard negative) 放入负样本集合再继续进行训练。困难样本对于提高特征的鲁棒性是有着很大帮助的，其原因再于利用一些困难样本进行有针对性的学习，可以让网络具有辨别困难任务的能力，从而能够在再识别过程中有更准确的判断。因此，如何发现困难样本和如何利用困难样本都是可以研究的点。例如说，可以在损失函数设计的时候，针对错分样本赋予更大的权重；又或者说提高困难样本参与网络训练的频率。

(2) 对于减少对于数据标注依赖的问题，除了本文所提出的利用伪标签以及聚类方法之外，还可以继续研究的是使用迁移学习来解决这个问题。然而对于行人再识别来说，域迁移的效果会受到很多因素的影响，比如图片背景、光照。如何在域迁移的过程中去除这些干扰因素的影响也同样是一个值得研究的问题。在此，一个可以进行尝试的方法在于自动对图片中的行人区域进行检测，从而移除背景部分再进行域迁移。

(3) 结果的再排序。在检索任务中，如何通过对结果的再排序来提升检索的准确率也是一个重要的研究点。行人再识别任务可以看成是一个检索的任务，进行再排序符合其应用场景并且能够提升其检索的鲁棒性。例如说，可以首先通过查询图片找到类似的结果，然后根据返回结果可以定位到源视频片段中，然后再在原视频片段中进行更为细致的查找。

(4) 对于现实生活中的行人再识别任务来说，庞大的数据处理量也对行人再识别的检索效率提出了更高的要求。查询检索的时间是根据特征尺寸和图库大小而显着增加的。就目前来说，在行人再识别领域，研究人员们并没有开展太多的着力于检索效率工作，而检索效率也是一个不可忽视的问题。然而，令人欣喜的是，针对

该点的研究可以从图像检索领域去借鉴一些方法，例如引入哈希特征表示，将其结合行人再识别任务的特征作出一些改进并应用到行人再识别上。

致　谢

似乎只在转眼间，紧张而艰苦的博士求学历程已即将结束。在我的博士论文完成之际，谨向在我攻读博士学位期间给我过我指导、帮助和关心的老师、同学和家人致以衷心的谢意。

首先衷心感谢我的导师白光一和唐振民教授。感谢白光一教授多年来对我的辛勤培育和悉心关怀，以及对我学术研究方向的指引，让我少走很多弯路，并对我生活上给与了充分的关系和支持。唐老师温文尔雅、而又不乏严厉，有幸成为唐老师的学生，是我一生的幸运。唐老师渊博的理论知识、敏锐的科学洞察力和积极的生活态度潜移默化的影响着我，他严谨的治学作风和执着的敬业精神是和时刻刻的鞭策着我。

感谢计算机学院张珊珊教授、蔡云飞副教授、吕建勇老师，悉尼科技大学的 Jian Zhang 老师以及澳大利亚国立大学的 Fatih Porikli 教授、Salman Khan 老师在学习和科研中给予的帮助。

感谢同门师兄弟诸葛程晨、刘家银、李祥瑞、刘华峰、张倚萌、姚凌翔、殷庆泽、孙泽人、罗皓楠、王毅、丁雨华、陈涛、江希若等对我的帮助，共同的学习生活，使我们结下了深厚的友谊，同时我也要感谢我的好友陈龙涛、成健、车继鲁等人在生活上对我的关切和帮助。

谨以本文献给我最亲爱的家人，是他们在背后持续无条件的关心和鼓励，才能支持我能够顺利完成学业！

参考文献

[1] Cai Q, Aggarwal J k. Tracking human motion in structured environments using a distributed-camera system[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPMAI), 1999, 21(11): 1241–1247.

[2] Wang X. Intelligent multi-camera video surveillance: A review[J]. Pattern Recognition Letters (PRL), 2013, 34(1): 3–19.

[3] Dollár P, Wojek C, Schiele B, et al. Pedestrian Detection: An Evaluation of the State of the Art[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPMAI), 2012, 34(4): 743–761.

[4] Zheng L, Yang Y, Hauptmann A g. Person re-identification: Past, present and future[J]. arXiv preprint arXiv:1610.02984, 2016.

[5] Zajdel W, Zivkovic Z, Krose B. Keeping track of humans: Have I seen this person before?[C] // Proceedings of the IEEE International Conference on Robotics and Automation (ICRA). 2005: 2081–2086.

[6] Gheissari N, Sebastian T b, Hartley R. Person reidentification using spatiotemporal appearance[C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2006: 1528–1535.

[7] Liu X, Song M, Zhao Q, et al. Attribute-restricted latent topic model for person re-identification[J]. Pattern Recognition (PR), 2012, 45(12): 4204–4213.

[8] Zheng L, Shen L, Tian L, et al. Scalable person re-identification: A benchmark[C] // Proceedings of the IEEE International Conference on Computer Vision (ICCV). 2015: 1116–1124.

[9] Deng W, Zheng L, Ye Q, et al. Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification[C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2018: 994–1003.

[10] Krizhevsky A, Sutskever I, Hinton G e. Imagenet classification with deep convolutional neural networks[C] // Proceedings of the Advances in Neural Information Processing Systems (NeurIPS). 2012: 1097–1105.

[11] Zheng L, Yang Y, Tian Q. SIFT Meets CNN: A Decade Survey of Instance Retrieval[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 2018, 40(5): 1224–1244.

[12] Liao S, Hu Y, Zhu X, et al. Person re-identification by local maximal occurrence representation and metric learning[C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2015 : 2197 – 2206.

[13] Li W, Zhao R, Xiao T, et al. Deepreid: Deep filter pairing neural network for person re-identification[C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2014 : 152 – 159.

[14] Zheng L, Bie Z, Sun Y, et al. Mars: A video benchmark for large-scale person re-identification[C] // Proceedings of the European Conference on Computer Vision (ECCV). 2016 : 868 – 884.

[15] Philbin J, Chum O, Isard M, et al. Object retrieval with large vocabularies and fast spatial matching[C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2007 : 1 – 8.

[16] Wang J, Zhang T, Sebe N, et al. A survey on learning to hash[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 2018, 40(4) : 769 – 790.

[17] Zheng W s, Gong S, Xiang T. Associating Groups of People[C] // British Machine Vision Conference (BMVC). 2009.

[18] Liu C, Gong S, Loy C c, et al. Person re-identification: What features are important?[C] // Proceedings of the European Conference on Computer Vision (ECCV). 2012 : 391 – 401.

[19] Zhang L, Xiang T, Gong S. Learning a discriminative null space for person re-identification[C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016 : 1239 – 1248.

[20] Ma X, Zhu X, Gong S, et al. Person re-identification by unsupervised video matching[J]. Pattern Recognition (PR), 2017, 65 : 197 – 210.

[21] Li M, Zhu X, Gong S. Unsupervised person re-identification by deep learning tracklet association[C] // Proceedings of the European Conference on Computer Vision (ECCV). 2018 : 737 – 753.

[22] Dong X, Yu S-i, Weng X, et al. Supervision-by-Registration: An unsupervised approach to improve the precision of facial landmark detectors[C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2018 : 360 – 368.

[23] Zhong Z, Zheng L, Zheng Z, et al. Camstyle: a novel data augmentation method for person re-identification[J]. IEEE Transactions on Image Processing (TIP), 2019, 28(3) : 1176 – 1190.

[24] Fan H, Zheng L, Yan C, et al. Unsupervised person re-identification: Clustering and

fine-tuning[J]. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 2018, 14(4): 83.

[25] Wu Y, Lin Y, Dong X, et al. Exploit the unknown gradually: One-shot video-based person re-identification by stepwise learning[C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2018: 5177–5186.

[26] Wei L, Zhang S, Gao W, et al. Person transfer gan to bridge domain gap for person re-identification[C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2018: 79–88.

[27] Sun Y, Zheng L, Yang Y, et al. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline)[C] // Proceedings of the European Conference on Computer Vision (ECCV). 2018: 480–496.

[28] Xie L, Wang J, Wei Z, et al. Disturblabel: Regularizing cnn on the loss layer[C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016: 4753–4762.

[29] Su C, Li J, Zhang S, et al. Pose-driven Deep Convolutional Model for Person Re-identification[C] // Proceedings of the IEEE International Conference on Computer Vision (ICCV). 2017: 3960–3969.

[30] Zhu X, Jing X-y, Wu F, et al. Distance learning by treating negative samples differently and exploiting impostors with symmetric triplet constraint for person re-identification[C] // Proceedings of the IEEE International Conference on Multimedia and Expo (ICME). 2016: 1–6.

[31] Gray D, Hai T. Viewpoint Invariant Pedestrian Recognition with an Ensemble of Localized Features[C] // Proceedings of the European Conference on Computer Vision (ECCV). 2008: 262–275.

[32] Farenzena M, Bazzani L, Perina A, et al. Person re-identification by symmetry-driven accumulation of local features[C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2010: 2360–2367.

[33] Prosser B j, Zheng W-s, Gong S, et al. Person re-identification by support vector ranking.[C] // British Machine Vision Conference (BMVC): Vol 2. 2010: 6.

[34] Zheng W-s, Gong S, Xiang T. Reidentification by relative distance comparison[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 2013, 35(3): 653–668.

[35] Ma A j, Yuen P c, Li J. Domain transfer support vector ranking for person re-identification without target camera label information[C] // Proceedings of the

IEEE International Conference on Computer Vision (ICCV). 2013: 3567–3574.

[36] Mignon A, Jurie F. PCCA: A new approach for distance learning from sparse pairwise constraints[C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2012: 2666–2672.

[37] Zheng W-s, Li X, Xiang T, et al. Partial person re-identification[C] // Proceedings of the IEEE International Conference on Computer Vision (ICCV). 2015: 4678–4686.

[38] Zhao R, Ouyang W, Wang X. Unsupervised salience learning for person re-identification[C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2013: 3586–3593.

[39] Li Z, Chang S, Liang F, et al. Learning locally-adaptive decision functions for person verification[C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2013: 3610–3617.

[40] Chen D, Yuan Z, Chen B, et al. Similarity learning with spatial constraints for person re-identification[C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016: 1268–1277.

[41] Zhao R, Ouyang W, Wang X. Person re-identification by salience matching[C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2013: 2528–2535.

[42] Zhao R, Ouyang W, Wang X. Learning mid-level filters for person re-identification[C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2014: 144–151.

[43] Shen Y, Lin W, Yan J, et al. Person re-identification with correspondence structure learning[C] // Proceedings of the IEEE International Conference on Computer Vision (ICCV). 2015: 3200–3208.

[44] Das A, Chakraborty A, Roy-chowdhury A k. Consistent re-identification in a camera network[C] // Proceedings of the European Conference on Computer Vision (ECCV). 2014: 330–345.

[45] Bazzani L, Cristani M, Perina A, et al. Multiple-shot person re-identification by hpe signature[C] // Proceedings of the IEEE International Conference on Pattern Recognition (ICPR). 2010: 1413–1416.

[46] Zhou X, Cui N, Li Z, et al. Hierarchical gaussianization for image classification[C] // Proceedings of the IEEE International Conference on Computer Vision (ICCV). 2009: 1971–1977.

[47] Chen D, Yuan Z, Hua G, et al. Similarity learning on an explicit polynomial kernel

feature map for person re-identification[C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2015 : 1565 – 1573.

[48] Pedagadi S, Orwell J, Velastin S, et al. Local fisher discriminant analysis for pedestrian re-identification[C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2013 : 3318 – 3325.

[49] Liu X, Song M, Tao D, et al. Semi-supervised coupled dictionary learning for person re-identification[C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2014 : 3550 – 3557.

[50] Yang Y, Yang J, Yan J, et al. Salient Color Names for Person Re-identification[C] // Proceedings of the European Conference on Computer Vision (ECCV). 2014 : 536 – 551.

[51] Zhang Y, Li B, Lu H, et al. Sample-specific svm learning for person re-identification[C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016 : 1278 – 1287.

[52] Van de weijer J, Schmid C, Verbeek J, et al. Learning color names for real-world applications[J]. IEEE Transactions on Image Processing (TIP), 2009, 18(7) : 1512 – 1523.

[53] Matsukawa T, Okabe T, Suzuki E, et al. Hierarchical gaussian descriptor for person re-identification[C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016 : 1363 – 1372.

[54] Layne R, Hospedales T m, Gong S, et al. Person re-identification by attributes.[C] // British Machine Vision Conference (BMVC) : Vol 2. 2012 : 8.

[55] Su C, Yang F, Zhang S, et al. Multi-task learning with low rank attribute embedding for person re-identification[C] // Proceedings of the IEEE International Conference on Computer Vision (ICCV). 2015 : 3739 – 3747.

[56] Shi Z, Hospedales T m, Xiang T. Transferring a semantic representation for person re-identification and search[C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2015 : 4184 – 4193.

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

[58] Yi D, Lei Z, Liao S, et al. Deep metric learning for person re-identification[C] // Proceedings of the IEEE International Conference on Pattern Recognition (ICPR). 2014 : 34 – 39.

[59] Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C] // Proceedings of the IEEE Conference on

Computer Vision and Pattern Recognition (CVPR). 2014: 580–587.

[60] Wu L, Shen C, van den Hengel A. Deep linear discriminant analysis on fisher networks: A hybrid architecture for person re-identification[J]. Pattern Recognition (PR), 2017, 65: 238–250.

[61] Perronnin F, Sánchez J, Mensink T. Improving the fisher kernel for large-scale image classification[C] // Proceedings of the European Conference on Computer Vision (ECCV). 2010: 143–156.

[62] Wu S, Chen Y-c, Li X, et al. An enhanced deep feature representation for person re-identification[C] // Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV). 2016: 1–8.

[63] Yang L, Jin R. Distance metric learning: A comprehensive survey[J]. Michigan State Universiy, 2006, 2(2): 4.

[64] Xing E p, Jordan M i, Russell S j, et al. Distance metric learning with application to clustering with side-information[C] // Proceedings of the Advances in Neural Information Processing Systems (NeurlPS). 2003: 521–528.

[65] Koestinger M, Hirzer M, Wohlhart P, et al. Large scale metric learning from equivalence constraints[C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2012: 2288–2295.

[66] Weinberger K q, Saul L k. Distance metric learning for large margin nearest neighbor classification[J]. Journal of Machine Learning Research (JMLR), 2009, 10(Feb): 207–244.

[67] Davis J v, Kulis B, Jain P, et al. Information-theoretic metric learning[C] // Proceedings of the International Conference on Machine Learning (ICML). 2007: 209–216.

[68] Hirzer M, Roth P m, Köstinger M, et al. Relaxed pairwise learned metric for person re-identification[C] // Proceedings of the European Conference on Computer Vision (ECCV). 2012: 780–793.

[69] Liao S, Li S z. Efficient psd constrained asymmetric metric learning for person re-identification[C] // Proceedings of the IEEE International Conference on Computer Vision (ICCV). 2015: 3685–3693.

[70] Yang Y, Liao S, Lei Z, et al. Large scale similarity learning using similar pairs for person verification[C] // Proceedings of the AAAI Conference on Artificial Intelligence (AAAI). 2016: 3655–3661.

[71] Mika S, Ratsch G, Weston J, et al. Fisher discriminant analysis with kernels[C] // Neural Networks for Signal Processing (NNSP). 1999: 41–48.

[72] Xiong F, Gou M, Camps O, et al. Person re-identification using kernel-based metric learning methods[C] // Proceedings of the European Conference on Computer Vision (ECCV). 2014: 1–16.

[73] Liu X, Wang H, Wu Y, et al. An ensemble color model for human re-identification[C] // Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV). 2015: 868–875.

[74] Zhu X, Jing X-y, You X, et al. Video-based person re-identification by simultaneously learning intra-video and inter-video distance metrics[J]. IEEE Transactions on Image Processing (TIP), 2018, 27(11): 5683–5695.

[75] Radenović F, Tolias G, Chum O. CNN image retrieval learns from BoW: Unsupervised fine-tuning with hard examples[C] // Proceedings of the European Conference on Computer Vision (ECCV). 2016: 3–20.

[76] Schroff F, Kalenichenko D, Philbin J. Facenet: A unified embedding for face recognition and clustering[C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2015: 815–823.

[77] Ahmed E, Jones M, Marks T k. An improved deep learning architecture for person re-identification[C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2015: 3908–3916.

[78] Wu L, Shen C, Hengel A v d. Personnet: Person re-identification with deep convolutional neural networks[J]. arXiv preprint arXiv:1601.07255, 2016.

[79] Varior R r, Shuai B, Lu J, et al. A siamese long short-term memory architecture for human re-identification[C] // Proceedings of the European Conference on Computer Vision (ECCV). 2016: 135–153.

[80] Varior R r, Haloi M, Wang G. Gated siamese convolutional neural network architecture for human re-identification[C] // Proceedings of the European Conference on Computer Vision (ECCV). 2016: 791–808.

[81] Liu H, Feng J, Qi M, et al. End-to-end comparative attention networks for person re-identification[J]. IEEE Transactions on Image Processing (TIP), 2017, 26(7): 3492–3506.

[82] Cheng D, Gong Y, Zhou S, et al. Person re-identification by multi-channel parts-based cnn with improved triplet loss function[C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016: 1335–1344.

[83] Su C, Zhang S, Xing J, et al. Deep attributes driven multi-camera person re-identification[C] // Proceedings of the European Conference on Computer Vision (ECCV). 2016: 475–491.

[84] Hadsell R, Chopra S, Lecun Y. Dimensionality reduction by learning an invariant mapping[C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2006 : 1735 – 1742.

[85] Xiao T, Li H, Ouyang W, et al. Learning deep feature representations with domain guided dropout for person re-identification[C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016 : 1249 – 1258.

[86] Zheng L, Zhang H, Sun S, et al. Person re-identification in the wild[C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017 : 1367 – 1376.

[87] Zhang S, Yang M, Cour T, et al. Query specific fusion for image retrieval[C] // Proceedings of the European Conference on Computer Vision (ECCV). 2012 : 660 – 673.

[88] Tolias G, Jégou H. Visual query expansion with or without geometry: refining local descriptors by feature aggregation[J]. Pattern recognition (PR), 2014, 47(10) : 3466 – 3476.

[89] Chum O, Philbin J, Sivic J, et al. Total Recall: Automatic Query Expansion with a Generative Feature Model for Object Retrieval[C] // Proceedings of the IEEE International Conference on Computer Vision (ICCV). 2007 : 1 – 8.

[90] Arandjelović R, Zisserman A. Three things everyone should know to improve object retrieval[C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2012 : 2911 – 2918.

[91] Shen X, Lin Z, Brandt J, et al. Object retrieval and localization with spatially-constrained similarity measure and k-nn re-ranking[C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2012 : 3013 – 3020.

[92] Bai S, Bai X. Sparse Contextual Activation for Efficient Visual Re-ranking[J]. IEEE Transactions on Image Processing (TIP), 2016, 25(3) : 1 – 1.

[93] Jegou H, Harzallah H, Schmid C. A contextual dissimilarity measure for accurate and efficient image search[C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2007 : 1 – 8.

[94] Qin D, Gammeter S, Bossard L, et al. Hello neighbor: Accurate object retrieval with k-reciprocal nearest neighbors[C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2011 : 777 – 784.

[95] Zhong Z, Zheng L, Cao D, et al. Re-ranking person re-identification with k-reciprocal encoding[C] // Proceedings of the IEEE Conference on Computer Vision

and Pattern Recognition (CVPR). 2017: 1318–1327.

[96] Garcia J, Martinel N, Micheloni C, et al. Person Re-Identification Ranking Optimisation by Discriminant Context Information Analysis[C] // Proceedings of the IEEE International Conference on Computer Vision (ICCV). 2015.

[97] Nguyen V, Ngo T d, Nguyen K m t t, et al. Re-ranking for person re-identification[C] // Proceedings of the International Conference on Soft Computing and Pattern Recognition (SoCPaR). 2013: 304–308.

[98] Ma A j, Li P. Query based adaptive re-ranking for person re-identification[C] // Proceedings of the Asian Conference on Computer Vision (ACCV). 2014: 397–412.

[99] Liu C, Chen C l, Gong S, et al. POP: Person Re-Identification Post-Rank Optimisation[C] // Proceedings of the IEEE International Conference on Computer Vision (ICCV). 2013.

[100] Zheng L, Wang S, Tian L, et al. Query-adaptive late fusion for image search and person re-identification[C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2015.

[101] Wei L, Yang W, Mukunoki M, et al. Common-near-neighbor analysis for person re-identification[C] // Proceedings of the IEEE International Conference on Image Processing (ICIP). 2012: 1621–1624.

[102] Garcia J, Martinel N, Gardel A, et al. Discriminant Context Information Analysis for Post-Ranking Person Re-Identification[J]. IEEE Transactions on Image Processing (TIP), 2017, 26(4): 1650–1665.

[103] Leng Q, Hu R, Chao L, et al. Person re-identification with content and context re-ranking[J]. Multimedia Tools and Applications (MTA), 2015, 74(17): 6989–7014.

[104] Ye M, Chen J, Leng Q, et al. Coupled-view based ranking optimization for person re-identification[C] // Proceedings of the International Conference on Multimedia Modeling (MMM). 2015: 105–117.

[105] Ye M, Chao L, Yi Y, et al. Person Re-identification via Ranking Aggregation of Similarity Pulling and Dissimilarity Pushing[J]. IEEE Transactions on Multimedia (TMM), 2016, 18(12): 2553–2566.

[106] Wang H, Gong S, Zhu X, et al. Human-in-the-Loop Person Re-identification[C] // Proceedings of the European Conference on Computer Vision (ECCV). 2016.

[107] Bai S, Bai X, Tian Q. Scalable person re-identification on supervised smoothed manifold[C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017: 2530–2539.

[108] Gray D, Brennan S, Tao H. Evaluating appearance models for recognition, reacquisition, and tracking[C] // Proceedings of the IEEE International Workshop on Performance Evaluation for Tracking and Surveillance (PETS) : Vol 3. 2007 : 1 – 7.

[109] Loy C c, Xiang T, Gong S. Multi-camera activity correlation analysis[C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2009 : 1988 – 1995.

[110] Cheng D s, Cristani M, Stoppa M, et al. Custom pictorial structures for re-identification.[C] // British Machine Vision Conference (BMVC) : Vol 1. 2011 : 6.

[111] Hirzer M, Beleznai C, Roth P m, et al. Person re-identification by descriptive and discriminative classification[C] // Proceedings of the Scandinavian Conference on Image Analysis (SCIA). 2011 : 91 – 102.

[112] Martinel N, Micheloni C. Re-identify people in wide area camera network[C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 2012 : 31 – 36.

[113] Li W, Zhao R, Wang X. Human reidentification with transferred metric learning[C] // Proceedings of the Asian Conference on Computer Vision (ACCV). 2012 : 31 – 44.

[114] Li W, Wang X. Locally aligned feature transforms across views[C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2013 : 3594 – 3601.

[115] Roth P m, Hirzer M, Köstinger M, et al. Mahalanobis distance learning for person re-identification[G] // Person re-identification. [S.l.] : Springer, 2014 : 247 – 267.

[116] Zheng Z, Zheng L, Yang Y. Unlabeled samples generated by gan improve the person re-identification baseline in vitro[C] // Proceedings of the IEEE International Conference on Computer Vision (ICCV). 2017 : 3774 – 3782.

[117] Ess A, Leibe B, Van gool L. Depth and appearance for mobile scene analysis[C] // Proceedings of the IEEE International Conference on Computer Vision (ICCV). 2007 : 1 – 8.

[118] Baltieri D, Vezzani R, Cucchiara R. 3dpes: 3d people dataset for surveillance and forensics[C] // Proceedings of the Joint ACM workshop on Human Gesture and Behavior Understanding (J-HGBU). 2011 : 59 – 64.

[119] Wang T, Gong S, Zhu X, et al. Person re-identification by video ranking[C] // Proceedings of the European Conference on Computer Vision (ECCV). 2014 : 688 – 703.

[120] Felzenszwalb P f, Girshick R b, Mcallester D, et al. Object detection with discrim-

inatively trained part-based models[J]. IEEE transactions on pattern analysis and machine intelligence (TPAMI), 2010, 32(9): 1627–1645.

[121] Ristani E, Solera F, Zou R, et al. Performance measures and a data set for multi-target, multi-camera tracking[C] //Proceedings of the European Conference on Computer Vision (ECCV). 2016: 17–35.

[122] Huang W, Hu R, Liang C, et al. Camera network based person re-identification by leveraging spatial-temporal constraint and multiple cameras relations[C] //Proceedings of the International Conference on Multimedia Modeling (MMM). 2016: 174–186.

[123] Singh K k, Lee Y j. Hide-and-Seek: Forcing a Network to be Meticulous for Weakly-Supervised Object and Action Localization[C] //Proceedings of the IEEE International Conference on Computer Vision (ICCV). 2017: 3544–3553.

[124] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016: 770–778.

[125] Hansen L k, Salamon P. Neural network ensembles[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 1990, 12(10): 993–1001.

[126] Sollich P, Krogh A. Learning with ensembles: How overfitting can be useful[C] //Proceedings of the Advances in Neural Information Processing Systems (NeurIPS). 1996: 190–196.

[127] Opitz D, Maclin R. Popular ensemble methods: An empirical study[J]. Journal of artificial intelligence research (JAIR), 1999, 11: 169–198.

[128] Zheng Z, Zheng L, Yang Y. Pedestrian alignment network for large-scale person re-identification[J]. IEEE Transactions on Circuits and Systems for Video Technology (TCSVT), 2018.

[129] Jaderberg M, Simonyan K, Zisserman A, et al. Spatial transformer networks[C] //Proceedings of the Advances in Neural Information Processing Systems (NeurIPS). 2015: 2017–2025.

[130] Wang F, Zuo W, Lin L, et al. Joint Learning of Single-Image and Cross-Image Representations for Person Re-Identification[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016: 1288–1296.

[131] Chen Y, Zhu X, Gong S, et al. Person re-identification by deep learning multi-scale representations[J]. Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017: 2590–2600.

[132] Wei L, Zhang S, Yao H, et al. Glad: global-local-alignment descriptor for pedestrian

retrieval[C] // Proceedings of the ACM International Conference on Multimedia (ACM-MM). 2017: 420 – 428.

[133] Li D, Chen X, Zhang Z, et al. Learning deep context-aware features over body and latent parts for person re-identification[C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017: 384 – 393.

[134] Fu J, Zheng H, Mei T. Look Closer to See Better: Recurrent Attention Convolutional Neural Network for Fine-Grained Image Recognition[C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017: 4438 – 4446.

[135] Vedaldi A, Lenc K. Matconvnet: Convolutional neural networks for matlab[C] // Proceedings of the ACM International Conference on Multimedia (ACM-MM). 2015: 689 – 692.

[136] Geng M, Wang Y, Xiang T, et al. Deep transfer learning for person re-identification[J]. ArXiv preprint arXiv:1611.05244, 2016.

[137] Ustinova E, Ganin Y, Lempitsky V. Multi-region bilinear convolutional neural networks for person re-identification[C] // Proceedings of the IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). 2017: 1 – 6.

[138] Liu J, Zha Z-j, Tian Q, et al. Multi-scale triplet cnn for person re-identification[C] // Proceedings of the ACM International Conference on Multimedia (ACM-MM). 2016: 192 – 196.

[139] Lin Y, Zheng L, Zheng Z, et al. Improving person re-identification by attribute and identity learning[J]. arXiv preprint arXiv:1703.07220, 2017.

[140] Hermans A, Beyer L, Leibe B. In Defense of the Triplet Loss for Person Re-Identification[J]. arXiv preprint arXiv:1703.07737, 2017.

[141] Li W, Zhu X, Gong S. Person Re-Identification by Deep Joint Learning of Multi-Loss Classification[C] // Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI). 2017: 2194 – 2200.

[142] Chen Y, Zhu X, Gong S. Person Re-Identification by Deep Learning Multi-Scale Representations[C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017: 2590 – 2600.

[143] Barbosa I b, Cristani M, Caputo B, et al. Looking beyond appearances: Synthetic training data for deep cnns in re-identification[J]. ArXiv preprint, 2017.

[144] Li W, Zhu X, Gong S. Harmonious attention network for person re-identification[C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2018: 2285 – 2294.

[145] Zheng L, Huang Y, Lu H, et al. Pose invariant embedding for deep person re-identification[J]. arXiv preprint arXiv:1701.07732, 2017.

[146] Zheng Z, Zheng L, Yang Y. A Discriminatively Learned CNN Embedding for Person Re-identification[J]. ACM Transactions on Multimedia Computing, Communications, and Applications (ACM-TOMM), 2017.

[147] Sun Y, Zheng L, Deng W, et al. SVDNet for Pedestrian Retrieval[C] // Proceedings of the IEEE International Conference on Computer Vision (ICCV). 2017: 3800–3808.

[148] Xiao T, Li S, Wang B, et al. Joint detection and identification feature learning for person search[C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017: 3415–3424.

[149] Goodfellow I, Pouget-abadie J, Mirza M, et al. Generative adversarial nets[C] // Proceedings of the Advances in Neural Information Processing Systems (NeurlPS). 2014: 2672–2680.

[150] Odena A. Semi-supervised learning with generative adversarial networks[J]. arXiv preprint arXiv:1606.01583, 2016.

[151] Salimans T, Goodfellow I, Zaremba W, et al. Improved techniques for training gans[C] // Proceedings of the Advances in Neural Information Processing Systems (NeurlPS). 2016: 2234–2242.

[152] Huang Y, Xu J, Wu Q, et al. Multi-Pseudo Regularized Label for Generated Data in Person Re-Identification[J]. IEEE Transactions on Image Processing (TIP), 2019, 28(3): 1391–1403.

[153] Szegedy C, Vanhoucke V, Ioffe S, et al. Rethinking the inception architecture for computer vision[C] // Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR). 2016: 2818–2826.

[154] Lee D-h. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks[C] // Proceedings of the International Conference on Machine Learning Workshop (ICMLW): Vol 3. 2013: 2.

[155] Mirza M, Osindero S. Conditional generative adversarial nets[J]. arXiv preprint arXiv:1411.1784, 2014.

[156] Zhu J-y, Park T, Isola P, et al. Unpaired image-to-image translation using cycle-consistent adversarial networks[J]. Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017: 2223–2232.

[157] Radford A, Metz L, Chintala S. Unsupervised representation learning with deep convolutional generative adversarial networks[J]. arXiv preprint arXiv:1511.06434,

2015.

[158] Arjovsky M, Chintala S, Bottou L. Wasserstein gan[J]. arXiv preprint arXiv:1701.07875, 2017.

[159] Gulrajani I, Ahmed F, Arjovsky M, et al. Improved training of wasserstein gans[C] // Proceedings of the Advances in Neural Information Processing Systems (NeurlPS). 2017: 5769 – 5779.

[160] Zhang C, Bengio S, Hardt M, et al. Understanding deep learning requires rethinking generalization[C] // Proceedings of the International Conference on Learning Representations (ICLR). 2017: 1 – 15.

[161] Wan L, Zeiler M, Zhang S, et al. Regularization of neural networks using dropconnect[C] // Proceedings of the International Conference on Machine Learning (ICML). 2013: 1058 – 1066.

[162] Ba J, Frey B. Adaptive dropout for training deep neural networks[C] // Proceedings of the Advances in Neural Information Processing Systems (NeurlPS). 2013: 3084 – 3092.

[163] Zeiler M d, Fergus R. Stochastic pooling for regularization of deep convolutional neural networks[C] // Proceedings of the International Conference on Learning Representations (ICLR). 2013: 1 – 9.

[164] Kang G, Dong X, Zheng L, et al. Patchshuffle regularization[J]. arXiv preprint arXiv:1707.07103, 2017.

[165] Qian X, Fu Y, Xiang T, et al. Pose-normalized image generation for person reidentification[C] // Proceedings of the European Conference on Computer Vision (ECCV). 2018: 650 – 667.

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

[167] Plaut D c, others. Experiments on Learning by Back Propagation.[J], 1986.

[168] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv:1409.1556, 2014.

[169] Zhong Z, Zheng L, Kang G, et al. Random erasing data augmentation[J]. arXiv preprint arXiv:1708.04896, 2017.

[170] Wang X, Shrivastava A, Gupta A. A-fast-rcnn: Hard positive generation via adversary for object detection[C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017: 2606 – 2615.

[171] Girshick R. Fast r-cnn[C] // Proceedings of the IEEE International Conference on

Computer Vision (ICCV). 2015 : 1440 – 1448.

[172] Srivastava N, Hinton G, Krizhevsky A, et al. Dropout: a simple way to prevent neural networks from overfitting[J]. The Journal of Machine Learning Research (JMLR), 2014, 15(1) : 1929 – 1958.

[173] Ioffe S, Szegedy C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift[C] // Proceedings of the International Conference on Machine Learning (ICML). 2015 : 448 – 456.

[174] Sukhbaatar S, Bruna J, Paluri M, et al. Training convolutional networks with noisy labels[J]. arXiv preprint arXiv:1406.2080, 2014.

[175] Jarrett K, Kavukcuoglu K, Lecun Y, et al. What is the best multi-stage architecture for object recognition?[C] // Proceedings of the IEEE international conference on computer vision (ICCV). 2009 : 2146 – 2153.

[176] Goodfellow I j, Warde-farley D, Mirza M, et al. Maxout networks[J]. arXiv preprint arXiv:1302.4389, 2013.

[177] Glorot X, Bordes A, Bengio Y. Deep sparse rectifier neural networks[C] // Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS). 2011 : 315 – 323.

[178] Liang X, Hu Z, Zhang H, et al. Recurrent topic-transition gan for visual paragraph generation[C] // Proceedings of the IEEE International Conference on Computer Vision (ICCV). 2017 : 3362 – 3371.

[179] Gurumurthy S, Kiran sarvadevabhatla R, Venkatesh babu R. Deligan: Generative adversarial networks for diverse and limited data[C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017 : 166 – 174.

[180] Marchesi M. Megapixel size image creation using generative adversarial networks[J]. arXiv preprint arXiv:1706.00082, 2017.

[181] Wen Y, Zhang K, Li Z, et al. A discriminative feature learning approach for deep face recognition[C] // Proceedings of the European Conference on Computer Vision (ECCV). 2016 : 499 – 515.

[182] Zhao L, Li X, Wang J, et al. Deeply-learned part-aligned representations for person re-identification[J]. arXiv preprint arXiv:1707.07256, 2017.

[183] Zhou S, Wang J, Wang J, et al. Point to set similarity based deep feature learning for person re-identification[C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017 : 3741 – 3750.

[184] Lin J, Ren L, Lu J, et al. Consistent-aware deep learning for person re-identification in a camera network[C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017: 5771–5780.

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

[186] Zheng Z, Zheng L, Yang Y. A Discriminatively Learned CNN Embedding for Person Reidentification[J]. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 2017, 14(1): 13.

[187] Schumann A, Stiefelhagen R. Person re-identification by deep learning attribute-complementary information[C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 2017.

[188] Ding G, Khan S, Tang Z, et al. Feature mask network for person re-identification[J]. Pattern Recognition Letters (PRL), 2019.

[189] Zhou Q, Fan H, Su H, et al. Weighted Bilinear Coding over Salient Body Parts for Person Re-identification[J]. arXiv preprint arXiv:1803.08580, 2018.

[190] Song C, Huang Y, Ouyang W, et al. Mask-guided contrastive attention model for person re-identification[C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2018: 1179–1188.

[191] Suh Y, Wang J, Tang S, et al. Part-aligned bilinear representations for person re-identification[C] // Proceedings of the European Conference on Computer Vision (ECCV). 2018: 402–419.

[192] Lisanti G, Masi I, Bagdanov A d, et al. Person re-identification by iterative re-weighted sparse ranking[J]. IEEE transactions on pattern analysis and machine intelligence (TPAMI), 2015, 37(8): 1629–1642.

[193] Wang H, Gong S, Xiang T. Unsupervised learning of generative topic saliency for person re-identification[C] // British Machine Vision Conference (BMVC). 2014.

[194] Kodirov E, Xiang T, Gong S. Dictionary learning with iterative laplacian regularisation for unsupervised person re-identification.[C] // British Machine Vision Conference (BMVC): Vol 3. 2015: 8.

[195] Ma B, Su Y, Jurie F. Bicov: a novel image representation for person re-identification and face verification[C] // British Machine Vision Conference (BMVC). 2012: 11.

[196] Ma B, Su Y, Jurie F. Local descriptors encoded by fisher vectors for person re-identification[C] // Proceedings of the European Conference on Computer Vision

(ECCV). 2012 : 413 – 422.

[197] Tzeng E, Hoffman J, Darrell T, et al. Simultaneous deep transfer across domains and tasks[C] // Proceedings of the IEEE International Conference on Computer Vision (ICCV). 2015 : 4068 – 4076.

[198] Long M, Cao Y, Wang J, et al. Learning transferable features with deep adaptation networks[C] // Proceedings of the International Conference on Machine Learning (ICML). 2015 : 97 – 105.

[199] Sun B, Saenko K. Deep coral: Correlation alignment for deep domain adaptation[C] // Proceedings of the European Conference on Computer Vision (ECCV). 2016 : 443 – 450.

[200] Ganin Y, Lempitsky V. Unsupervised Domain Adaptation by Backpropagation[C] // Proceedings of the International Conference on Machine Learning (ICML). 2015 : 1180 – 1189.

[201] Sun B, Feng J, Saenko K. Return of frustratingly easy domain adaptation[C] // Proceedings of the AAAI Conference on Artificial Intelligence (AAAI). 2016.

[202] Tzeng E, Hoffman J, Zhang N, et al. Deep domain confusion: Maximizing for domain invariance[J]. arXiv preprint arXiv:1412.3474, 2014.

[203] Zhong Z, Zheng L, Li S, et al. Generalizing a person retrieval model hetero-and homogeneously[C] // Proceedings of the European Conference on Computer Vision (ECCV). 2018 : 172 – 188.

[204] Ganin Y, Ustinova E, Ajakan H, et al. Domain-adversarial training of neural networks[J]. The Journal of Machine Learning Research (JMLR), 2016, 17(1) : 2096 – 2030.

[205] Peng P, Xiang T, Wang Y, et al. Unsupervised cross-dataset transfer learning for person re-identification[C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016 : 1306 – 1315.

[206] Wang J, Zhu X, Gong S, et al. Transferable joint attribute-identity deep learning for unsupervised person re-identification[C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2018 : 2275 – 2284.

[207] Lin Y, Dong X, Zheng L, et al. A bottom-up clustering approach to unsupervised person re-identification[C] // Proceedings of the AAAI Conference on Artificial Intelligence (AAAI) : Vol 2. 2019.

[208] Gower J c. A comparison of some methods of cluster analysis[J]. Biometrics, 1967 : 623 – 637.

[209] Deng J, Dong W, Socher R, et al. Imagenet: A large-scale hierarchical image database[C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2009: 248 – 255.

[210] Maaten L v d, Hinton G. Visualizing data using t-SNE[J]. Journal of machine learning research (JMLR), 2008, 9(Nov): 2579 – 2605.

[211] Ye M, Ma A j, Zheng L, et al. Dynamic label graph matching for unsupervised video re-identification[C] // Proceedings of the IEEE International Conference on Computer Vision (ICCV). 2017: 5142 – 5150.

[212] Liu Z, Wang D, Lu H. Stepwise metric promotion for unsupervised video person re-identification[C] // Proceedings of the IEEE International Conference on Computer Vision (ICCV). 2017: 2429 – 2438.

[213] Ye M, Lan X, Yuen P c. Robust anchor embedding for unsupervised video person re-identification in the wild[C] // Proceedings of the European Conference on Computer Vision (ECCV). 2018: 170 – 186.

[214] Chen Y, Zhu X, Gong S. Deep association learning for unsupervised video person re-identification[C] // British Machine Vision Conference (BMVC). 2018.

附 录

攻读博士学位期间发表的论文和出版著作情况：

[1] **Guodong Ding**, Shanshan Zhang, Salman Khan, Zhenmin Tang, Jian Zhang, Fatih Porikli, Feature Affinity based Pseudo Labeling for Semi-supervised Person Re-identification[J]. IEEE transactions on Multimedia. 2019, 21(11): 2891 - 2902. (中科院 SCI 二区，影响因子 3.977)

[2] **Guodong Ding**, Salman Khan, Zhenmin Tang, Fatih Porikli, Feature mask network for person re-identification[J]. Pattern Recognition Letters, 2019. https://doi.org/10.1016/j.patrec.2019.02.015. (中科院 SCI 三区，影响因子 1.954)

[3] **Guodong Ding**, Salman Khan, Zhenmin Tang, Dispersion based Clustering for Unsupervised Person Re-identification[C] // British Machine Vision Conference (BMVC), 2019.(EI 会议)

[4] **Guodong Ding**, Shanshan Zhang, Salman Khan, Zhenmin Tang, Center based Pseudo-labeling for Semi-supervised Person Re-identification[C] // IEEE International Conference on Multimedia and Expo Workshops (ICMEW), 2018.(EI 会议)

[5] **Guodong Ding**, Salman Khan, Zhenmin Tang, Jian Zhang, Fatih Porikli, Validity Guided Unsupervised Person Re-identification with Cluster Dispersion[J]. IEEE transcations on Image Processing. (审稿中，中科院 SCI 二区，影响因子 5.072)