中国科学技术大学

学士学位论文

中国科学技术大学

基于多模态数据的无监督行人重识别

作者姓名： 孙启

学科专业： 天文学

导师姓名： 李厚强教授 周文罡教授

完成时间： 二〇二二年五月二十四日

University of Science and Technology of China

A dissertation for bachelor's degree

Unsupervised Person Re-Identification with Multi-modal Data

Author: Qi Sun

Speciality: Astronomy

Supervisors: Prof. Houqiang Li, Prof. Wengang Zhou

Finished time: May 24, 2022

致　谢

在科大的本科生活逐渐到了尾声。这四年虽快，但却是我人生中改变最大的一段时光。这里，我要感谢一路上遇到的人，是他们带给我成长和蜕变。

首先，我要感谢我的导师李厚强教授和周文罡教授。在大三的暑假，两位老师给完全没有科研经验和背景知识的我进入 MCC 实验室的机会，也给予我宝贵的指导。非常有幸在研究生阶段，我还能继续跟随两位老师进行计算机视觉的研究。其次，我要感谢刘一衡师兄，他给了我非常细致、耐心的指导。从一开始和我讨论行人重识别领域的经典文章，再到寒假期间几乎每天和我开会，叮嘱我跑实验，以及和我讨论最终的实验方案，这些所有的情形我都历历在目。

在物院的三年，我也得到了很多老师和同学的关心和帮助。王俊贤教授指导了我近两年在活动星系核方面的研究。在他的帮助下，我不再只迷恋于天文中的“big picture”，而是沉下心来细致地进行数据处理和统计分析。我同样感谢薛永泉教授，他是我在科研和生活上的榜样。尽管没能在天文的道路上走下去，和天文的故事仍让我收益匪浅。

我要感谢我这四年来的室友，他们不仅给我带来了夜聊的笑声，也教会了我如何在集体中生活。我还要感谢初高中以来一直帮助我的朋友们。每次和他们聊起对生活、学习的感悟，时常引起共鸣，让我感到不再孤单，倍感振奋。

我要感谢我的女友苏航，她总能给我快乐。

最后，我的父母是我最要感谢的人，他们是我二十年来最坚实的后盾。疫情期间，平时很难见着面，只能电话交流。每天几分钟的聊天里，是家长里短，是生活里的趣事。而且，我总能从他们那里听到有用的建议：狠狠摔倒时，他们相信我非常棒；踌躇满志时，他们会告诉我前路漫漫；科研迷茫、无力时，他们总能让我不要钻牛角尖，学会健康，积极的生活。我的成长离不开他们的呵护和关爱。

目 录

中文内容摘要

行人重识别是在数据库中检索出跨摄像头下特定行人的计算机视觉技术。现有的行人重识别方法只是利用视觉信息来得到行人的表征，但是视觉信息很容易受到遮挡，模糊，行人衣物更换等因素影响，从而很难在实际应用中训练出一个可靠的模型。考虑到当今社会很多人都随身携带移动设备，行人的无线定位信息很容易通过基站定位或 WiFi 定位等方式获取。因此，基于弱场景标注的多模态行人重识别任务开始引起关注，其可以在仅知道相机位置的情况下利用行人手机的无线定位信号辅助行人重识别。现有方法使用图神经网络融合多模态信息，但其对多模态数据的利用不充分，性能有限。为了解决这个问题，我们设计了多模态聚类对比学习框架，其主要包含亲和度融合模块和多模态交替优化策略。亲和度融合模块综合考虑视觉信息和无线线索，得到视频之间稳健的距离度量；多模态交替优化策略利用视觉标签和无线标签交替训练编码器，使得编码器更容易收敛到最优模型。我们在 WP-ReID 和 Campus4K 数据集上做了充分的实验，验证了所提方法的有效性。同时，实验结果表明，我们的方法较现有方法有很大的性能优势，在两个数据集上的平均均值精度分别有 16.5% 和 9.8% 的性能提升。代码见 https://github.com/iamsunqi/wireless-contrast。

关键词：行人重识别；无监督学习；无线定位；对比学习；多模态

Abstract

Person re-identification is a computer vision technique that retrieves a specific person captured by another camera from the image gallery. Existing unsupervised person re-identification models rely on visual information to obtain the representation of pedestrians. Since visual clues are sensitive to occlusion, blur, clothes changing of pedestrians, etc., it is hard to train a reliable model for real-world applications. Considering many people carry their mobile equipment with them in our society, it is accessible to pedestrians' wireless positioning information from the base station or WiFi positioning. Therefore, the task, i.e. *person re-identification under weak scene labeling with multi-modal data* , attracts extensive attention from the community, in which setting, we use wireless positioning information from pedestrians' phones to assist person re-identification when only knowing the location of cameras. The existing method uses graph neural networks (GNN) to integrate multi-modal information but this method can not fully explore the multi-modal data, resulting in a limited performance. To solve this, we propose a Multi-Modal Clustering Contrastive Framework (MMCCF) , in which affinity fusion module and multimodal alternating optimization strategy are used. Affinity fusion module considers both visual information and wireless clues to get robust metrics on video pairs, while multi-modal alternating optimization strategy uses visual labels and wireless labels to train the encoder in an alternating manner, making the encoder much easier to converge into an optimal model. We conduct extensive experiments in WP-ReID and Campus4K dataset to validate the effectiveness of our proposed method. In the meanwhile, our proposed method outperforms all other unsupervised models by a large margin, with an improvement of 16.5% and 9.8% on mAP in two datasets, respectively. The code is available at https://github.com/iamsunqi/wireless-contrast.

Key Words: Person Re-Identification; Unsupervised Learning; Wireless Positioning; Contrastive Learning; Multi-Modal

第一章　绪论

第一节　研究背景

计算机视觉是人工智能的一个领域，指让计算机能够从图像，视频等输入中获得有意义的信息，从而赋予计算机发现、感知和理解的能力。相较于生物视觉，计算机视觉依赖于摄像机、数据、算法和计算，而不是通过视网膜和视觉神经来完成。经过训练，一个计算机视觉系统可以在适当条件下，短时间完成对成百上千个图片或视频的分析，并且能发现极小的缺陷或问题，这是生物视觉所做不到的。而行人重识别作为计算机视觉一个重要的子课题，近年来得到了学术界和工业界越来越多的关注。行人重识别技术旨在利用计算机视觉算法，检索出跨摄像头的特定行人的图片或视频。

行人重识别技术具有很高的应用价值和现实意义。通过行人重识别系统，可以借助部署的摄像头，迅速找到不同场景下的目标人物。智能安防领域中，警方可以利用重识别系统侦测可疑人员；智能商业中，商家可以对特定用户进行轨迹追踪，了解用户的兴趣特点，用以优化用户购物体验；在大型的娱乐场所，比如迪士尼公园中，就可以部署行人重识别系统来实现人物寻回和定位。如果走失儿童没有携带手环或其他 GPS 工具，那么则需要通过视觉系统来进行跟踪和定位；摄像头分辨率低，人脸只占图片的很小部分，人脸背对或者侧对摄像头，这些情况都会使人脸识别技术对此应用场景无用武之地。对行人的整体识别此时则可以起到关键的作用。

行人重识别技术有着重要的应用价值，但同时也面临着巨大的技术挑战。比如，摄像头拍摄的行人图片普遍存在遮挡现象，这会使得一些有区分力的部分无法被摄像头捕捉到，给行人个体的表征学习引入了很大的噪声。除此之外，摄像头拍摄照片的模糊现象、不同时间光照情况不同、行人姿态和衣物的变化等，都会降低视觉数据的可靠性，可能让行人重识别系统匹配到错误的行人。

本文首先梳理了基于深度学习的行人重识别领域的进展；然后针对弱场景标注下的无监督行人重识别问题，提出了一个基于对比学习的多模态无监督行人重识别算法，在大规模数据集上证明了算法的有效性和可行性，为进一步发展跨领域大数据的综合应用提供了指导。

第二节　基于深度学习的行人重识别研究现状

行人重识别任务旨在利用一个行人的图片，在跨摄像头的数据库中检索到同样身份的行人图片。它被广泛认为是图像检索的子问题。在前深度学习时代，研究者主要关注于如何设计好行人的手工特征以及特征之间的距离度量。而神经网络作为一种强大的特征提取器，深度学习开始作为行人重识别的主流方法[1][2][3]。基于深度学习的行人重识别的典型思路就是利用训练好的编码器(如残差网络[4]) 将图片编码成视觉表征，再利用简单的欧式距离度量两个图片的相似度，就可以取得很好的性能。而且，基于深度学习的行人重识别系统可以实现端到端的训练，使任务变得简单优雅。基于深度学习的行人重识别具有很高的研究价值以及现实应用意义，大量研究工作发表在 CVPR, ICCV, ECCV, NeurIPS, TPAMI 等顶级会议和期刊上。

一、基于图片的行人重识别

上面提到，行人重识别一个重要任务就是学习如何提取好的特征，即特征学习。在行人重识别技术发展的早期，常常会对一个人物图片抽取一个整体的特征。为了解决不同图片的整体特征可能不对应的问题，学习局部/区域的特征也是一个很好的思路。Sun 等[5] 提出了一个基于部分卷积神经网络的基线模型，多个分类器用于判别水平划分的局域特征，达到了 SOTA 的性能。

在深度学习方法出现之前，基于度量学习的行人重识别也得到了广泛的研究，旨在学习特征之间的好的距离度量（比如马氏距离）。但由于特征表示学习的兴起，度量学习逐步被损失函数的设计所取代。下面介绍主要的损失函数：

身份损失. 它把行人重识别问题看作图片分类问题，每一个行人代表一个类别。对于一个“图片，标签”组 $(\boldsymbol{x}_i, y_i)$，由 $\boldsymbol{x}_i$ 得到 y_i 的概率（利用 softmax 函数）被表示成 $p(y_i|\boldsymbol{x}_i)$。因此，身份损失可以由交叉熵损失函数计算而来

$$L_{id} = -\frac{1}{n}\sum_{i}^{n} \log p(y_i|\boldsymbol{x}_i), \tag{1.1}$$

其中 n 是一个小批量中含有的图片数。

对比损失. 它用于优化两个图片之间的距离关系，

$$L_{con} = (1-\delta_{ij})\{\max(0, \rho - d_{ij})\}^2 + \delta_{ij} d_{ij}^2, \tag{1.2}$$

其中 d_{ij} 代表两个图片特征 $\boldsymbol{x}_i$ 和 $\boldsymbol{x}_j$ 之间的欧式距离，δ_{ij} 是指示函数，ρ 是一个阈值变量。如果两个输入的样本属于同一个身份，则最小化它们之间的距离；如

果不属于同一个样本，则尽量让两者之间的距离大于 ρ。对比损失常常结合身份损失来共同提升性能。

三元组损失. 它把行人重识别模型的训练过程看成是检索排序问题。它基本的想法是正样本对之间的距离一定要小于负样本对之间的距离。一个典型的三元组包括一个锚点样本 $\boldsymbol{x}_i$，一个正样本 $\boldsymbol{x}_j$，一个负样本 $\boldsymbol{x}_k$，则三元组损失定义为:

$$L_{tri}(i,j,k) = \max(\rho + d_{ij} - d_{ik}, 0). \tag{1.3}$$

如果直接优化三元组损失，问题是显然的。训练样本中很大一部分都是容易的三元组，导致训练效果不佳。为了解决这个问题，基本的想法是找出富含信息的三元组。Hermans 等[6] 在一个训练批量中在线发掘最难的正样本和负样本，这样可以训练出一个具有判别力的模型。同时使用三元组损失和身份损失是现在监督行人重识别最流行的解决方式。

二、基于视频的行人重识别

在基于视频的行人重识别任务中，数据集由视频序列组成，采用与图像行人重识别一样的方法不能充分利用视频的时域信息。为了补充卷积神经网络仅能利用空间维度的缺陷，学者开始利用循环神经网络及其变种长短期记忆用于视频序列的建模。

McLaughlin 等[7] 将输入信息分为表观特征和光流信息，在卷积神经网络基础上加上循环神经网络使之可以处理视频序列，在循环神经网络层加入时域池化使之可以处理任意长度的视频。Liu 等[8] 提出时空线索整合模块，将时间特征和空间特征相互促进并融合成视频稳健的时空特征。该工作在 iLIDS-VID[9]、PRID-2011[10]、MARS[11] 等大规模视频数据集上都取得了很好的性能。

三、基于对比学习的行人重识别

虽然有监督情形下的行人重识别模型性能达到了很高的水平，但是在实际应用中，给每个行人都标注上身份是非常消耗精力和资源的。所以无监督行人重识别近年来收到学术界和工业界很多的关注[12]，比如：Li 等[13] 提出了一个无监督视频重识别框架, 同时探索同一个摄像头下的视频关联和跨摄像头的视频关联；Ge 等[14] 提出了一个自步对比学习框架。从 2020 年 MoCo[15] 和 SimCLR[16] 提出开始，计算机视觉领域掀起了一股基于对比学习的自监督表征学习的狂潮，它们展示了在很多视觉任务上，用很多无标签的数据做训练甚至可以超过很多

有监督算法。作为计算机视觉的重要应用，无监督行人重识别也受到这股浪潮的影响，出现了很多优秀的工作，这些算法的性能直逼有监督训练的效果。

对比学习是实现自监督学习的重要方式，它通过学习两个个体是否相似来编码得到表征。对比学习的核心思想就是锚点 $\boldsymbol{x}$ 和正样本之间距离远远大于锚点和负样本之间的距离，即

$$s(f(\boldsymbol{x}), f(\boldsymbol{x}^+)) >> s(f(\boldsymbol{x}), f(\boldsymbol{x}^-)). \tag{1.4}$$

为优化锚点与其正负样本之间的关系，使用内积作为距离函数，并构造 softmax 分类器，设计得到 infoNCE loss[17], 即

$$L_N = E_X[\log \frac{\exp(f(\boldsymbol{x}) \cdot f(\boldsymbol{x}^+))}{\exp(f(\boldsymbol{x}) \cdot f(\boldsymbol{x}^+)) + \sum_{j=1}^{N-1} \exp(f(\boldsymbol{x}) \cdot f(\boldsymbol{x}_j))}]. \tag{1.5}$$

对比学习中构建正负样本很灵活，也是问题的核心。对序列数据而言，CPC[18] 提出将这段序列 t 之后的输入作为正样本，从其他序列中随机采出样本作为负样本。Wu 等[17] 提出记忆银行 (memory bank)，将之前模型产生的特征都存起来，计算损失时可以之间索引，模型更新完后将当前特征重新存入记忆银行 (memory bank)。MoCo[15] 使用队列数据结构来存储特征，并采用动量更新的编码器。SimCLR[16] 使用不同的数据增强来构造正负样本。

Ge 等[19] 首先提出了一个基于混合记忆的自步对比学习框架和统一对比损失，充分利用了源域、目标域的数据; 进而 Dai 等[20] 提出了建立在记忆银行基础上 (memory bank) 更简洁的 ClusterNCE loss。其中 $\{\boldsymbol{f}_i\}_{i=1}^N$ 存储着不同类的特征，N 是所有类的个数，ClusterNCE loss 可以写成:

$$L_q = -\log \frac{\exp(\langle \boldsymbol{q}, \boldsymbol{f}_+ \rangle / \tau)}{\sum_{i=1}^N \exp(\langle \boldsymbol{q}, \boldsymbol{f_i} \rangle / \tau)}, \tag{1.6}$$

其中 $\boldsymbol{q}_i$ 是属于 $\boldsymbol{f}_i$ 第 i 个类的索引特征，τ 是控制相似空间 (similarity space) 的温度变量；$\langle \boldsymbol{q}, \boldsymbol{f}_i \rangle$ 衡量索引向量和类特征的相似度。其中类特征在训练的过程中得到在线更新。在 Market1501,MSMT17 和 DukeMTMC-ReID 都取得了 SOTA 的性能。Chen 等[21] 通过拉近不同摄像头内同一个类的特征来学习得到更有判别力的模型。

对比学习成为了无监督行人重识别一个主流的解决方案。

四、跨模态的行人重识别

单纯的视觉信息在行人重识别任务上取得了巨大的成就。但考虑到实际应用中，行人通常会携带手机，这些 GPS 信号等无线信息可以为行人重识别提供

更加可靠的线索，从而提升行人重识别系统的稳健性。

Liu 等[22] 介绍了全场景标注下无线信息和视觉信息融合的算法。它提出了一种循环上下文传播单元 (Recurrent Context Propogation Module, RCPM) 来循环迭代视频之间的无线相似度和视觉相似度。Liu 等[23] 考虑到全场景标注的昂贵性，引入了弱场景标注下的无线信息和视觉信息融合的算法。它利用基于自适应聚类的多模态数据关联来探索视频和无线信号之间可能的关联性，利用多模态图神经网络来学习视频之间的相似度。

第三节　行人重识别数据集

不管是监督算法还是无监督算法，都需要在大规模的行人重识别数据集上面进行训练，并且评估效果。大规模数据集往往可以推动技术的发展，就类似于 ImageNet 数据集之于视觉识别技术的作用。幸运的是，现在有很多大规模的数据集可供使用。这些常见的数据集的统计信息见表1.1。

表 1.1　大规模行人重识别数据集统计

数据集	发布年份	行人数量	相机数目	图片数量	标注方式	视频
CUHK03[1]	14’	1360	6	13164	手工/DPM[24]	否
Market1501[25]	15’	1501	6	32668	手工/DPM[24]	否
MARS[26]	15’	2161	6	1067516	DPM[24]	是
DukeMTMC-ReID[27]	17’	1812	8	36441	手工	否
DukeMTMC-VideoReID[28]	18’	1401	8	815420	手工	是
MSMT17[29]	18’	4101	15	126441	Faster R-CNN[30]	否
SYSU-MM01[31]	18’	4901	6	45863	-	否
Campus4K[32]	19’	1567	6	521309	YOLO[33]	是
WPReID[22]	20’	79	6	106578	SiamRPN[34]	是

注：这里列出 9 种大规模的行人重识别数据集，其中 WPReID[22] 还具有和视觉数据关联的无线数据，也是本文中将要实验的数据集。SYSU-MM01[31] 数据集中的 45863 图片包括 30071 张 RGB 图片和 15792 张红外图片。

一、图片数据集

CUHK03[1] 是第一个大规模的行人重识别数据集。它包含在 6 个摄像机下 1360 个行人的图片，包含 13164 张图片。这个数据集同时包括了手工标注和自动检测的检测框。其中，1160 个行人的图片被用于训练，200 个行人的图片用于

测试。

Market1501[25] 是图片行人重识别问题中一个基准数据集。它有清华大学校园内的一个超市旁的 6 个摄像头采集。它一共包括 1501 个行人和 32668 张图片。平均每个行人在每个摄像机下有 3.6 个图片。行人检测框采用 DPM[24] 检测器。其中，750 个行人的图片用于训练，其余 751 个行人图片被用于评估测试。测试集包括 19732 张图片，其中 3368 张图片作为查询集。

DukeMTMC-ReID[27] 是从 DukeMTMC 跟踪数据集采样得到的一个字瞬间。这些行人照片是从 8 个不重叠的摄像头手动裁剪而来。它包含 1404 个行人和 14183 个行人的照片，每个行人至少在两个摄像头下出现过。其中，训练集包括 702 个行人的 16522 张图片。测试集包括 702 个行人的 17661 张图片，其中 2228 张图片作为查询集。

MSMT17[29] 是图片行人重识别的大规模数据集。它从由 15 个摄像机拍摄的，12 段时间间隔内采集的视频中采集而来。这些视频总共 180 小时长，有很强的光照改变。作为一个图片行人重识别数据集，它包括 4101 个行人的 126441 张图片。

二、视频数据集

MARS[26] 是一个大规模视频数据集。这些视频由 6 个摄像机采集而来。1261 个行人中每个人至少在两个摄像头下出现过。光照条件的改变和图片质量的不一在这个数据集中很常见，这给基于视频行人重识别问题带来了很大的挑战。这些视频是通过 DPM 检测器和 GMMCP 跟踪器得到。这个数据包含 631 个行人的训练集和 630 个行人的测试集。

DukeMTMC-VideoReID[28] 是 DukeMTMC 跟踪数据集的一个子集。它包含 8 个摄像头下，1812 个行人的 4832 个视频序列。平均来看，每个视频序列包含连续的 168 帧，其中每一个行人的检测框都是手工裁剪并标注的。

Campus4K[32] 是第一个 4K 分辨率的视频数据集。它是在科大校园内 6 个不重叠的摄像头采集而来，所有的时空信息都有相应记录。高分辨率的视频甚至可以辨识人脸，这将人脸识别任务和重识别任务结合起来。它包含 1567 个行人的 3849 个视频序列（521309 帧）。其中，695 个行人的 1843 视频序列作为训练集（243998 帧），查询集包括 695 个行人的 695 个视频序列（97914 帧），库集包括 872 个行人的 1311 个视频序列（179397 帧）。

三、多模态数据集

WPReID[22] 是一个包含无线大数据的视频行人重识别数据集。其中无线大数据由行人携带手机的无线定位功能所提供。行人每走过监控摄像头，摄像机会记录行人的视频数据；行人运动时的手机无线定位信息也会被同事记录。它包含 79 个行人的 868 个视频（106578 帧）。该数据集的提出，扩展了传统行人重识别的应用场景，提出了行人重识别问题的新设定。本文中正是基于该数据集来进行无线数据和视觉数据融合算法的训练和测试。

SYSU-MM01[31] 是一个为跨模态行人重识别任务设计的数据集，包含红外和可见光数据。SYSU-MM01 利用 6 个摄像头（4 个可见光摄像头、2 个红外摄像头）拍摄了 491 个行人的可见光、红外照片。可见光相机在明亮的环境下工作，而红外摄像头在黑暗的环境下工作。数据集中总共有 30071 个可见光照片和 15792 个红外照片。

第四节　行人重识别模型评价指标

行人重识别算法的性能主要通过累计匹配特征表（Cumulative Match Characteristic, CMC）和均值平均精度来衡量（mean Average Precision, mAP）。下面，介绍两种指标的具体形式。

一、累计匹配特征

在测试中，我们常用查询集中的图片来到库集中检索非同一摄像头下的图片。对于每一个查询图片（query），算法会按照库集图片（gallery）和查询图片的距离，依次将所有库集图片进行排序。CMC rank-k 准确率：

$$Acc_k = \begin{cases} 1, \text{如果前}k\text{个库样本中包含查询个体;} \\ 0, \text{其他情况.} \end{cases} \tag{1.7}$$

最终的 CMC 曲线是将所有的查询图片的 rank-k 准确率平均而来。

二、均值平均精度

均值平均精度被广泛使用作为目标检测，图像检索等领域的测评指标。当然，它也是行人重识别领域的一个重要测评指标。

前面提到，利用 CMC 曲线可以评估行人重识别算法的性能。CMC 曲线展

示了查询图片出现在不同大小的候选列表中出现的概率。但是，这种评估手段只在只有一个正确匹配（库集中）的情形下有效，如下图1.1（a)。如果存在多个正确匹配，那么 CMC 曲线则显得没有十分合理，因为检索的“召回率”没有得到考虑。在下图中，图1.1（b）和图1.1（c）的 CMC 曲线（rank-1）都是 1，但是很显然这两个检索列表的质量是完全不一样的。

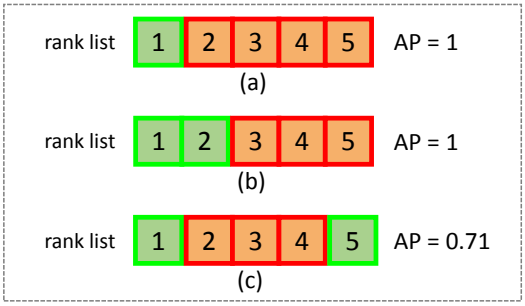

图 1.1　例子：CMC 和 AP 对检索结果评估的不一致性。绿色框是正确匹配，红色框是错误匹配。这三个检索列表中，CMC 曲线都是 1；但是 AP 分别为 1，1 和 0.71。

在 Market1501 数据集内[26]，对于每个查询图片，平均有 14.8 个跨摄像头的正确匹配。因此，Zheng 等[26] 首次提出用 mAP 来评价整体性能。对于每个查询图片，平均精度定义为准确率-召回率曲线（Precision-Recall Curve）下方的面积。然后，所有查询图片的平均精度的平均值，即 mAP，就可以被计算出来。它同时考虑到了算法的准确性和召回率，为算法提供了一个较为全面的评价。

第二章　多模态聚类对比学习框架

图2.1是本文提出的**多模态聚类对比学习框架**的整体网络架构。我们采取了迭代式的训练策略来优化编码器。为了更好的优化编码器，我们设计了**多模态交替优化策略**，其中每个训练周期分为两个阶段，第一个阶段仅使用视觉数据聚类得到伪标签进行监督训练；第二阶段融合无线信息，得到新的多模态伪标签进行训练。具体来说，初始化阶段，我们得到各视频的伪标签，形成记忆字典；训练过程中，利用查询图片的特征和记忆字典对比，更新编码器。首先，在第一节中，我们介绍问题的数学描述；第二节中介绍本文采用的对比学习基线模型；最后一节中，我们具体介绍**多模态聚类**生成无线伪标签的过程。

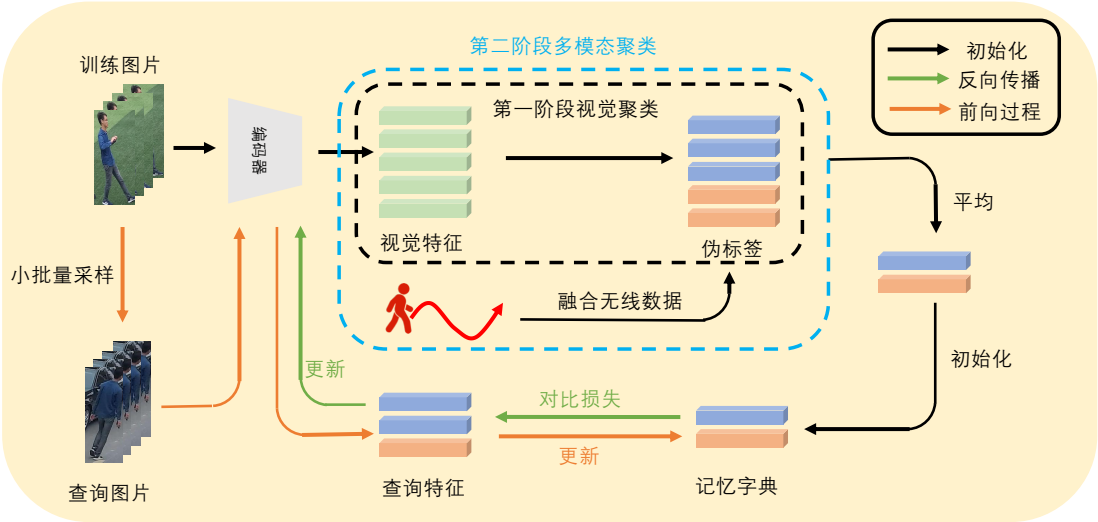

图 2.1　多模态聚类对比学习框架和多模态交替优化策略

第一节　问题描述

我们问题设定如图2.2所示：灰色背景是摄像系统所在的街区；蓝色扇形区域是摄像头监控区域；红色实线是行人轨迹；黑色圆形虚线是摄像头下的无线感知区域。行人每经过摄像头监控区域，即红色实线每与蓝色扇形区域有交集，便记录下一段视频序列。行人每经过一个无线感知区域，即红色实线每与黑色圆形区域有交集，便记录下一段无线片段。无监督的背景下，我们不知道行人的标签；弱场景标注的设定下，我们仅知道摄像头的位置，而不知道无线轨迹精确的时

空标注。这个任务，即弱场景标注下利用无线数据辅助行人重识别问题，由 Liu 等[23] 首次提出。

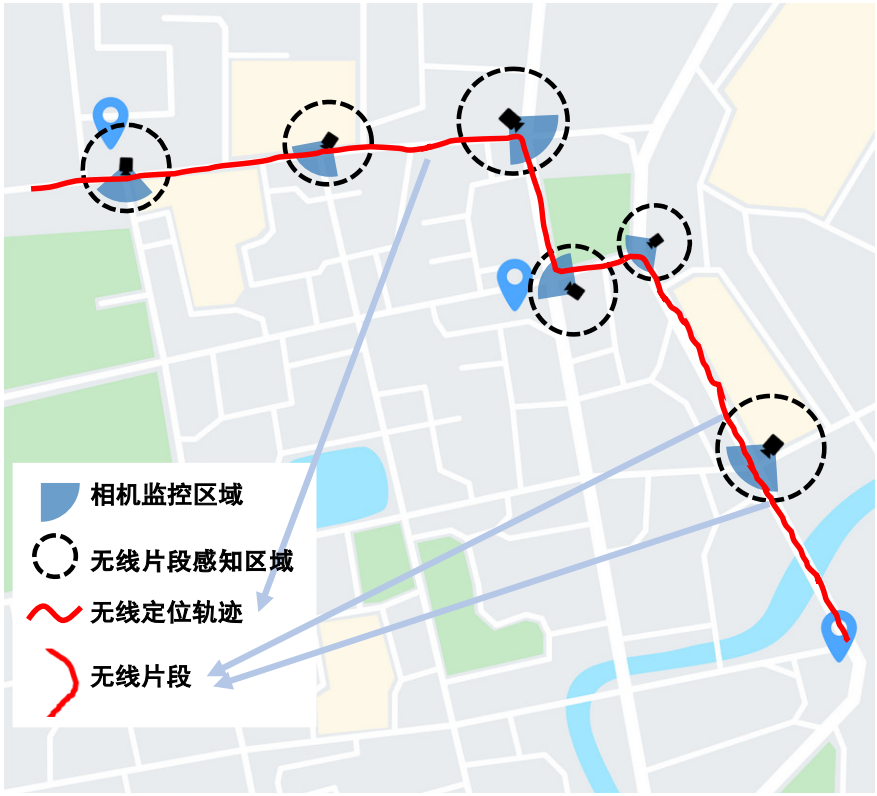

图 2.2　弱场景标注下的无监督重识别问题

所知很少的信息使得这个问题极具挑战性。但是，无线信息和视觉信息具有互补性和潜在的关联，这让问题的解决提供了潜在的可行性。具体来说，视觉数据是对行人细粒度的描述，但对遮挡，光照，角度变换，衣物更换等情形过于敏感；无线数据粗略地描述行人的时空信息，但以上的细节很难影响无线数据。另外，视频和对应无线片段有时间交集的特性为两者之间建立了可能的联系。

第二节　对比学习基线模型

我们的基线模型只采用图2.1中第一阶段视觉聚类，没有融合无线数据，这个模型很大程度参考了 Dai 等[20] 提出的类对比无监督行人重识别模型。

这里，简要说明一下用到的记号：训练集 $V=\{\boldsymbol{v}_i\}_{i=1}^{N}$ 由 N 个视频构成；编码器 $f_\theta(\cdot)$；$U=\{\mathbf{u}_1,\mathbf{u}_2,\cdots,\mathbf{u}_N\}$ 是对应的视频特征，其中 $\mathbf{u}_i=f_\theta(\mathbf{v}_i), i=1:N$。小批量的查询图片特征 $\mathbf{q}$ 在训练集 V 采样得到。

基线模型包括初始化和训练两个部分。其中初始化为模型提供记忆字典，同时为视频（图片）分配伪标签。训练则利用之前得到伪标签，和伪标签对应的记忆特征计算对比损失，即 ClusterNCE loss，用于训练编码器。以下会逐一介绍。

记忆初始化. 首先，我们使用编码器将图片编码成特征，再把视频内各帧特征进行平均得到视频层面的特征。然后，对视频层面的特征进行聚类得到每个视频对应的伪标签（类）。每个训练图片的伪标签由视频的伪标签决定。同时，我们把每个类的特征 $\{\boldsymbol{\phi}_1, \cdots, \boldsymbol{\phi}_K\}$ 存储在基于记忆的特征字典里，其中 K 是总类数。我们使用类内视频特征的平均值来初始化类的特征，即:

$$\boldsymbol{\phi}_k == \frac{1}{|H_k|} \sum_{\mathbf{u}_i \in H_k} \mathbf{u}_i, \tag{2.1}$$

其中 H_k 指属于第 k 个类的视频特征的集合。

伪标签的生成：最近邻关联. 在记忆的初始化中，我们采用最近邻关联作为视频特征的聚类方法。使用编码器可以得到每一个视频的特征 $U = \{\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_N\}$，最近邻关联（Nearest Neighbor Association, NNA）的方法可以用来联系跨摄像头的视频序列对，进而得到相应的伪标签。具体来说，对每个视频序列都可以找到跨摄像头的最近邻的视频序列。这里的距离度量是使用特征之间的余弦距离。如果这两个视频序列是**互为最近邻**，那么它们被视作同一个人，将被分配同一个伪标签。这里的最近邻关联是很强的条件，所以不是所有视频都被分配有伪标签。当然，如果不使用余弦距离，而采用其他的距离度量，只要能得到一个距离矩阵，就可以生成相应伪标签。

编码器训练和记忆更新. 在训练过程中，训练集每次会采样 P 个人物（同一伪标签视为一个人物），每个人物的 K 个图片。因此，一个小批量（mini-batch）中包含 $P \times K$ 个查询图片。如图2.3所示，对于一个查询图片 $\boldsymbol{q}$，在记忆字典中索引得到对应类特征，并加以更新类，即:

$$\boldsymbol{\phi}_+ \leftarrow m\boldsymbol{\phi}_+ + (1-m)\boldsymbol{q}. \tag{2.2}$$

同时，与对应类特征对比得到 ClusterNCE 损失，即

$$L_q = -\log \frac{\exp(\langle \boldsymbol{q}, \boldsymbol{f}_+ \rangle / \tau)}{\sum_{i=1}^{N} \exp(\langle \boldsymbol{q}, \boldsymbol{f_i} \rangle / \tau)}. \tag{2.3}$$

对比损失反向传播，更新编码器。其中 τ 是退火温度参数。

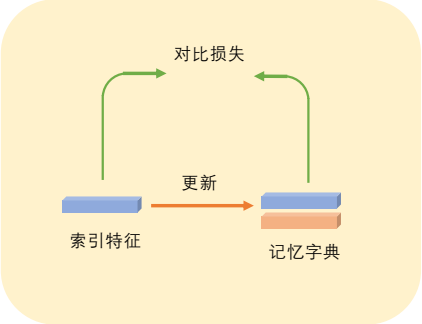

图 2.3　训练过程中，记忆更新和损失计算

第三节　多模态聚类

在基线模型中，编码器得到的视觉特征直接进行聚类得到视觉伪标签。而在多模态聚类中，如图2.4所示，我们会利用多模态数据关联模块[23]（Multi-Modal Data Association, MMDA）将无线数据和视觉信息联系起来，得到无线亲和度；进而在亲和度融合模块（Affinity Fusion Module, AFM）融合无线亲和度和视觉亲和度，重新最近邻关联聚类得到无线伪标签。在每个训练的迭代周期的第二个阶段，利用无线伪标签进行记忆字典初始化和伪标签的分配。以下将逐一详细介绍主要的模块。

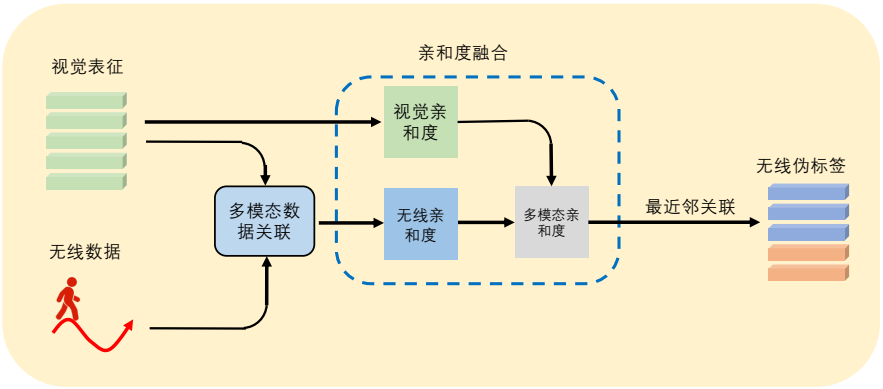

图 2.4　多模态聚类：多模态数据关联和亲和度融合模块

多模态数据关联：探索无线数据和视觉数据的可能关联。如图2.5所示，一条无线轨迹由多个无线片段组成，图示 T_m 和 $\{T_m^r\}_{r=1}^4$。在无线片段对应的起始时间段内，有很多对应的视频，图示 T_m^r 和 V_m^r。视频集合内的视频都有可能和无线片段或无线轨迹相对应，这就为无线数据和视觉数据建立了可能的联系。把这些

视频集合放在一起，再进行 K-means 聚类。前面，我们建立了无线片段和对应视频集之间较弱的时空关系。现在，我们假设每一个类都是一个独立的行人，就可以把每个类中视频和无线轨迹之间建立定量概率关系。如公式 (2.4) 所示，$Q(C_k)$ 函数返回在该类中，类中所有视频属于所有的无线片段的个数；R_m 代表无线轨迹总共包含的无线片段个数；因此，可以用 $P_{m,k}$ 来估计第 k 个类中视频属于这条无线轨迹的概率，见公式 (2.4)。如图2.5所示，比如在第 3 个类 $C_{m,3}$ 中，我们看到类中视频 1，2，3，4 分别属于 4 个无线片段；总无线片段数为 4；所以这里我们估计这个类中视频属于这条无线轨迹的概率是 $P_{m,3} = 4/4$。

$$P_{m,k} = \frac{Q(C_{m,k})}{R_m}. \tag{2.4}$$

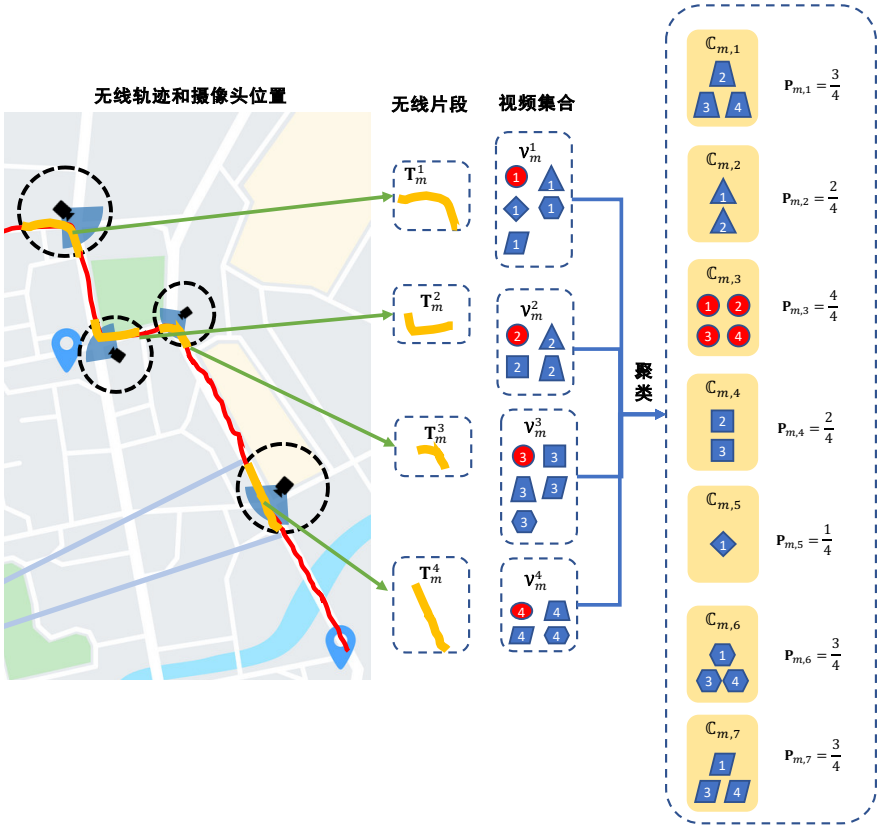

图 2.5　多模态数据关联

对关联无线轨迹 T_m 所有视频进行 K-means 聚类时，K 值 K_m 是重要的超参数。它由自适应估计得到，如公式（2.5）所示。我们的想法是用 K_m 估计无线轨迹 T_m 下的人数。$\sum_{r=1}^{R_m} |V_m^r|$ 是该无线轨迹下的总视频数。$\frac{\sum_{m=1}^{M} R_m}{M}$ 是平均每条无线轨迹的无线片段数。一个行人在监控场景下行走形成一条无线轨迹，每形成

一条无线轨迹同时有一条视频记录；所以 $\frac{\sum_{m=1}^{M} R_m}{M}$ 也可以估计一个行人平均视频数。所以将两者作比，$\frac{\sum_{r=1}^{R_m} |V_m^r| \times M}{\sum_{m=1}^{M} R_m}$ 就可以作为一个无线轨迹下行人个数的恰当估计。考虑到，在实际情况中，一个行人（无线轨迹）拥有的无线片段数不能准确反映一个行人生成的视频数。因此修正上述公式，修正系数 λ 控制 K_m 和估计行人个数的比值。

$$K_m = \lambda \frac{\sum_{r=1}^{R_m} |V_m^r| \times M}{\sum_{m=1}^{M} R_m}. \tag{2.5}$$

上面，我们建立了每条无线轨迹和其关联视频的定量概率关系。这样，我们进一步可以将每条无线轨迹下，与其关联视频之间的关系，即视频之间的亲和度，建立起来。如公式（2.6）所示，两个相同视频的亲和度定为 1；对于两个不同视频，如果它们存在于同一个类中，则它们的亲和度为这个类属于无线轨迹的概率；否则，两个视频的亲和度为 0。

$$S_{i,j,m} = \begin{cases} P_{m,k}, \text{如果 } i \neq j \text{ 且 } \exists\, \mathbf{C}_{m,k}, \{V_i, V_j\} \subset \mathbf{C}_{m,k}; \\ 1, \text{如果} i = j; \\ 0, \text{其他情况.} \end{cases} \tag{2.6}$$

亲和度融合模块利用每一个无线信号下，视频之间的亲和度 $\mathbf{S}_{i,j,m}$，对无线维度取平均，则得到（利用无线信息估计的）视频之间的亲和度 $\mathbf{A}_{i,j} = \sum_m \mathbf{S}_{i,j,m}/M$。

在实际大规模应用场景中，无线信号的数目也会非常多。这里直接使用所有无线通道矩阵的平均值就不甚合理，会让总无线亲和度矩阵过于平滑（over-smoothing），即使两个视频在某条无线信号上有较强的关联，所得到矩阵上相应数值也会非常接近于 0。所以，我们增加了过平滑去除处理，即对两个视频之间的亲和度，取前 15 个最大的无线通道的平均值作为视频之间的无线亲和度。

接下来，为了计算方便，我们采用距离来代替亲和度[①]。相应地，利用无线信息估计的视频之间的距离矩阵为 $\mathbf{D}_{i,j}^{w} = 1 - \mathbf{A}_{i,j}$。相应地，可以定义视觉信息得到的视频之间的距离矩阵 $\mathbf{D}_{i,j}^{v}$ [②]。综合两种亲和度（距离），我们定义总的距离矩阵为：$\mathbf{D}_{i,j}^{\tau} = \kappa \cdot \mathbf{D}_{i,j}^{w} + (1-\kappa) \cdot \mathbf{D}_{i,j}^{v}$。利用融合距离（亲和度）矩阵，我们可以进一步进行最近邻关联，从而给视频赋上新的伪标签，即无线伪标签。利用无线伪标签，就可以进行第二阶段的训练。

[①]视频之间亲和度和距离之和恒为 1；比如两个视频没有关联，那么它们的亲和度为 0，距离为 1。

[②]这里把具有相同视觉伪标签的视频间亲和度定义为 1（距离为 0），把不同视觉伪标签的视频的亲和度定义为 0（距离为 1）。

第三章　实验

本章在两个行人重识别数据集 WP-ReID[22] 和 Campus4K[32] 上评估本文提出的算法。

第一节　实验设置

本章用在 ImageNet 上预训练的 ResNet50[4] 作为编码器，其中最后的全连接层被移除。

初始化设置：多模态数据关联内，对于 WP-ReID 数据集和 Campus4K 数据集，λ 分别设置成 2 和 3。WP-ReID 数据集中无线数据的感知半径设置为 50m。亲和度融合模块内，无线亲和度矩阵和视觉亲和度矩阵的融合因子设置为 0.9。

训练设置：每个训练周期中，每一个视频序列中随机选取 60 帧（不满 60 帧的视频则选取全部帧），作为训练集。训练编码器时，使用 Adam[35] 优化器更新 100 个迭代周期，初始学习率为 0.0004，权重衰减为 5×10^{-4}。小批量设置为 256，每次采样 16 个行人（类）的 16 张图片。分别利用无线伪标签和视觉伪标签交替训练 40 个迭代周期，每次初始化过后（每个训练周期前）评估一次伪标签的调整后互信息（Adjusted Mutual Information, AMI）[①]和聚类个体占比，每隔 10 个训练周期，测试一次模型性能。对比学习中，记忆更新的动量为 0.2，温度设置为 0.05。本实验训练和测试均在 2 块 Nvidia 3090Ti 进行。

第二节　消融实验

一、伪标签的预测

视觉伪标签从视觉特征聚类得到，但很难区分部分视觉上很相近却不属于同一个人的“难样本”。无线信号则可以利用时空上的联系，将“难样本”和对应的属于同一个人的样本关联起来，赋予同一个伪标签。无线信号的鲁棒性为视觉信息提供了可能的补充。这里，我们使用调整后互信息量化聚类准确度。伪标签的准确度和无监督学习的性能直接相关。图3.1显示了训练的过程中两种聚类准确度的变化。如图3.1所示，在达到基本收敛之前（约 15 迭代周期），在两个数据

[①]通过计算真实标签和聚类预测标签的 AMI，可以用于估计聚类质量。具体见python 机器学习库 sklearn 手册。

集上视觉标签和无线标签的准确度都逐步提高。在 Campus4K 数据集上，无线聚类的准确度可以达到 97%，视觉聚类的准确度也可以达到近 90%；在 WP-ReID 数据集上，无线聚类和视觉聚类的准确度都可以达到 85% 左右。

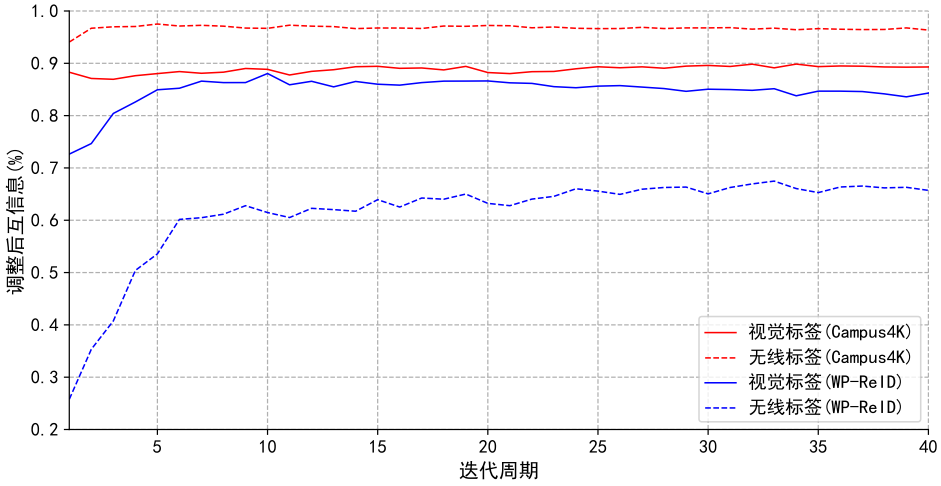

图 3.1　无线聚类和视觉聚类准确度在训练过程中的变化

一些难区分的图片没有被分配有无线标签。另外，由于不是所有视频都能有对应的无线信号，部分视频/图片没有被分配有无线标签。没有被分配伪标签的图片/视频不参与训练，所以未参与训练或者是未被分配有伪标签的图片/视频的比例对模型性能也起到至关重要的影响。图3.2展示了在训练过程中，两种聚类中没有伪标签图片/视频的比例。如图3.2所示，得益于迭代式的训练策略，在训练收敛前（约 10 个迭代周期），两个数据集上未被聚类的比例显著降低。特别地，在 Campus4K 数据集上，虽然无线伪标签在收敛后仍有约 30% 的图片/视频没有参与训练，但它有着 97% 以上的聚类准确度（图3.1所示）。

二、最近邻关联对基线模型的影响

最近邻关联是本文中实现聚类的重要方法，它利用 N 个视频之间的亲和度矩阵 $\boldsymbol{A} \in \mathbf{R}^{N \times N}$，将它们分别归类到 K 个类 $\{\boldsymbol{c}_k\}_{k=1}^{K}$ 中。它为基线模型和无线视觉亲和度融合模块都提供了非常准确的聚类方法。而准确的聚类为伪监督的训练提供了基础，给高性能提供了保障。在大部分无监督行人重识别的文献中[21][20][19]，都采用 DBSCAN[36] 或者 infomap[37] 的聚类方法，对图片进行聚类。相比以上两种聚类方法，最近邻关联是无参数的，避免了人为超参数设定的影响。表3.1展

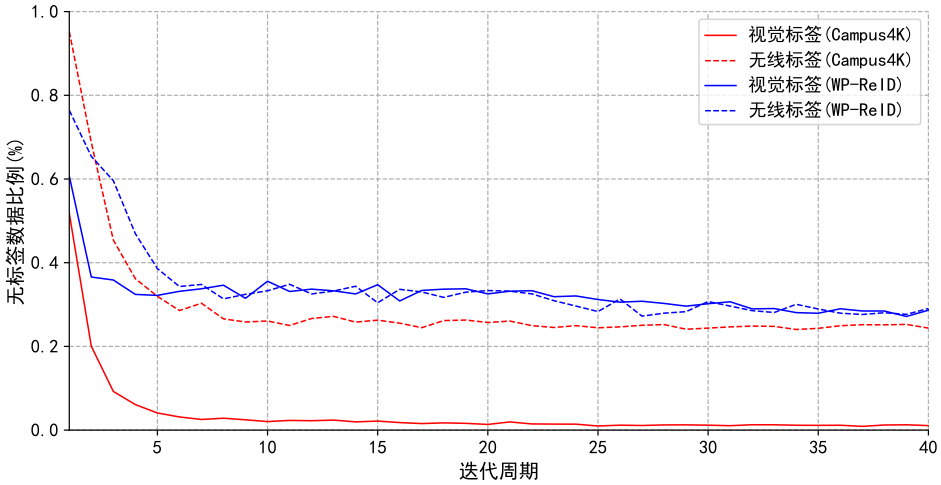

图 3.2　无线聚类和视觉聚类中未参与训练的样本占比在训练过程中的变化

示了采用不同聚类方法在基线模型上的性能差异，最近邻关联聚类取得了更好的性能。

表 3.1　最近邻关联和其他聚类方法的对比

数据集	WP-ReID				Campus4K			
方法	mAP	Rank@1	Rank@5	Rank@10	mAP	Rank@1	Rank@5	Rank@10
Baseline(infomap)	32.1	65.2	84.6	92.0	85.3	82.9	94.4	97.0
Baseline(NNA)	62.7	86.1	93.0	94.5	94.2	92.6	98.6	98.6

三、亲和度融合模块和多模态交替优化策略的影响

多模态数据关联把不同无线通道内视频之间的亲和度编码到一个无线亲和度矩阵中。亲和度融合模块（Affinity Fusion Module，AFM）将无线亲和度矩阵和视觉亲和度矩阵融合，再进行最近邻关联，得到新的无线伪标签。多模态交替优化策略采用在阶段一，先利用视觉标签训练一次编码器（基线模型），再在阶段二，利用无线伪标签训练一次编码器，交替训练。

我们考虑了我们方法的两种变体。一种简单的方法是不采用视觉伪标签训练编码器，只使用无线伪标签训练编码器。这意味着仅仅采用图2.4中阶段二来训练。另一种方法是直接利用不同无线下多模态数据关联得到的类，作为无线伪标签（Wireless Clustering, WC）训练编码器，替代亲和度融合模块得到的无线伪标签训练编码器。这两种方法在 WP-ReID 和 Campus4K 数据集上的性能都不如

我们的方法，甚至在 Campus4K 数据集上不如基线模型的表现，具体见表3.2。

表 3.2　直接无线聚类（WC）和亲和度融合（AFM）的对比

数据集	WP-ReID				Campus4K			
方法	mAP	Rank@1	Rank@5	Rank@10	mAP	Rank@1	Rank@5	Rank@10
Baseline(阶段一)	62.7	86.1	93.0	94.5	91.9	89.9	98.4	99.0
+WC(多模态交替优化)	70.2	87.1	93.5	95.5	81.0	75.5	94.9	97.1
+AFM(阶段二)	70.3	84.6	93.0	94.5	79.9	77.3	93.6	96.4
+AFM(多模态交替优化)	**73.5**	**89.6**	**94.5**	**97.0**	**94.2**	**92.6**	**99.3**	**99.4**

四、融合因子 κ 对性能的影响

总亲和矩阵由无线亲和矩阵和视觉亲和矩阵融合而来。其中控制两者比例的参数，即融合因子 κ，则是一个比较重要的超参数。图3.3中，我们探究了融合因子和模型性能的关系。

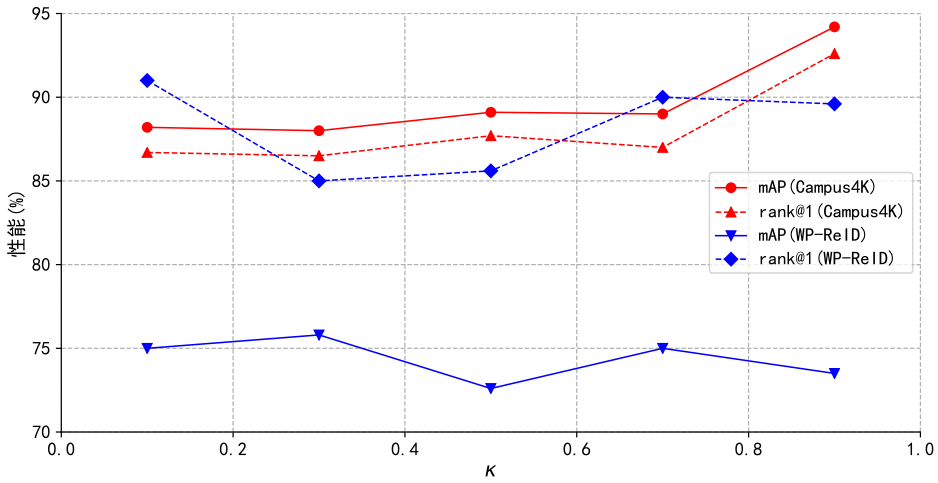

图 3.3　两个数据集上，融合因子 κ 对模型性能的影响

如图3.3所示, 在 WP-ReID 数据集上, 性能对融合因子不敏感, mAP 和 rank@1 分别在 0.1 和 0.3 处取得最大值；在 Campus4K 数据集上，可以看到性能随融合因子变大而递增，mAP 和 rank@1 均在 0.9 时取最大值。

因此，在结果中，我们统一采用 0.9 作为融合因子的值。

五、过平滑去除处理的影响

对于 Campus4K 数据集来说，包含着过多无线数据（80% 的视频都有无线信号)。为了解决在大规模无线信号的场景，我们对两个视频之间的亲和度，取前 15 个最大的无线通道的平均值作为视频之间的无线亲和度。表3.3展示了，在 Campus4K 数据集上，无线亲和度矩阵的过平滑去除处理（Oversmoothing Removal)，可以得到更好的融合亲和度，从而取得更好的聚类效果和模型性能。

表 3.3　过平滑去除 (Oversmoothing Removal) 在 Campus4K 数据集上对性能和聚类效果的影响

方法	AMI(%)	无标签比例 (%)	mAP	Rank@1	Rank@5	Rank@10
Baseline	-	-	91.9	89.9	98.4	99.0
+AFM(w/o OR)	89.9	78.8	88.6	86.4	97.4	98.8
+AFM(w OR)	96.4	24.3	**94.2**	**92.6**	**99.3**	**99.4**

第三节　与现有方法对比

表 3.4　所提方法和现有无监督方法在 WP-ReID 数据集对比

方法	参考文献	mAP	Rank@1	Rank@5	Rank@10
SpCL[19]	NeurIPS 20’	30.4	50.2	74.6	83.1
IICS[38]	CVPR 21’	30.4	50.2	74.6	83.1
MCDSCE[39]	CVPR 21’	30.4	50.2	74.6	83.1
ICE[21]	ICCV 21’	43.8	68.2	80.6	86.6
UMTF[23]	arXiv 21’	57.0	86.1	93.5	95.0
Baseline	我们的方法	62.7	86.1	93.0	94.5
Baseline+ AFM	我们的方法	**73.5**	**89.6**	**94.5**	**97.0**

本节在表3.4中将我们的方法和其他现有的无监督行人重识别方法在 WP-ReID 数据集上进行了对比。实验结果表明，我们的基线模型的 mAP 和 Rank-1 准确度已经大幅度超过现有纯视觉无监督方法,[21]；甚至超过了加上无线数据的 UMTF[23] 的性能。这表明我们对比学习基线模型的有效性。同时，我们的亲和度融合模块有效的整合了多模态信息，让 mAP 和 Rank-1 准确度又分别提升了 8.8% 和 3.5%。这表明引入无线轨迹来帮助无监督行人重识别是很有必要的，无

线信息和视觉数据可以相互补充，减少噪声的影响。

表 3.5　所提方法和现有无监督方法在 Campus4K 数据集对比

方法	参考文献	mAP	Rank@1	Rank@5	Rank@10
TASTR[32]	TMM 20’	84.4	82.7	94.6	-
SpCL[19]	NeurIPS 20’	54.8	48.8	77.6	85.4
IICS[38]	CVPR 21’	67	64.5	85.7	91.6
MCDSCE[39]	CVPR 21’	77.7	74.5	91.9	95.4
ICE[21]	ICCV 21’	67.1	64.4	84.5	90.0
UMTF[23]	arXiv 21’	66.3	91.0	96.0	97.0
Baseline	我们的方法	91.9	89.9	98.4	99.0
Baseline+AFM	我们的方法	**94.2**	**92.6**	**99.3**	**99.4**

表3.5将我们的方法和其他现有无监督行人重识别方法在 Campus4K 数据集上进行了对比。和在 WP-ReID 数据集上一样，我们基线模型和亲和度融合模块都取得了非常高的性能。

第四章　总结

本文结合无线数据和视觉数据，为解决弱场景标注下的无监督行人重识别问题，提出了一个多模态聚类对比学习框架。本文的主要贡献两点。第一，本文提出的多模态交替更新策略用视觉标签和无线标签交替训练编码器。第二，本文利用多模态数据关联寻找弱场景标注下无线数据和视觉数据可能的关联，并提出亲和度融合模块同时考虑多模态信息，从而达到了现有方法中最好的性能。

本文对无线大数据辅助行人重识别问题进行了初步探索，展示了不同质的多模态数据提升模型性能上巨大的潜力。但是本文中对无线数据约束方面的考虑远不够深入。另外，在实际应用部署中，无线数据的利用会面临更多工程上的挑战。这些都需要未来的工作去探索。

参 考 文 献

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

[2] YE M, SHEN J, LIN G, et al. Deep learning for person re-identification: A survey and outlook[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 2021.

[3] ZHENG L, YANG Y, HAUPTMANN A G. Person re-identification: Past, present and future[J/OL]. CoRR, 2016, abs/1610.02984. http://arxiv.org/abs/1610.02984.

[4] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]// In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016: 770-778.

[5] SUN Y, ZHENG L, YANG Y, et al. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline)[C]//In Proceedings of the European Conference on Computer Vision (ECCV). 2018: 480-496.

[6] HERMANS A, BEYER L, LEIBE B. In defense of the triplet loss for person re-identification[J/OL]. CoRR, 2017, abs/1703.07737. http://arxiv.org/abs/1703.07737.

[7] MCLAUGHLIN N, MARTINEZ DEL RINCON J, MILLER P. Recurrent convolutional network for video-based person re-identification[C]//In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016: 1325-1334.

[8] LIU Y, YUAN Z, ZHOU W, et al. Spatial and temporal mutual promotion for video-based person re-identification[C]//In Proceedings of the AAAI Conference on Artificial Intelligence (AAAI). 2019.

[9] WANG T, GONG S, ZHU X, et al. Person re-identification by video ranking[C]// In Proceedings of the European Conference on Computer Vision (ECCV). 2014: 688-703.

[10] HIRZER M, BELEZNAI C, ROTH P M, et al. Person re-identification by descriptive and discriminative classification[C]//In Proceedings of the Scandinavian

Conference on Image Analysis (SCIA). Springer, 2011: 91-102.

[11] ZHENG L, BIE Z, SUN Y, et al. Mars: A video benchmark for large-scale person re-identification[C]//In Proceedings of the European Conference on Computer Vision (ECCV). Springer, 2016: 868-884.

[12] LIN X, REN P, YEH C, et al. Unsupervised person re-identification: A systematic survey of challenges and solutions[J/OL]. CoRR, 2021, abs/2109.06057. https://arxiv.org/abs/2109.06057.

[13] LI M, ZHU X, GONG S. Unsupervised tracklet person re-identification[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 2019, 42(7): 1770-1782.

[14] GE Y, CHEN D, LI H. Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification[J]. In Proceedings of the International Conference on Learning Representations (ICLR), 2020.

[15] HE K, FAN H, WU Y, et al. Momentum contrast for unsupervised visual representation learning[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2020: 9726-9735.

[16] CHEN T, KORNBLITH S, NOROUZI M, et al. A simple framework for contrastive learning of visual representations[C/OL]//III H D, SINGH A. Proceedings of Machine Learning Research: volume 119 Proceedings of the 37th International Conference on Machine Learning. PMLR, 2020: 1597-1607. https://proceedings.mlr.press/v119/chen20j.html.

[17] WU J, YANG Y, LIU H, et al. Unsupervised graph association for person re-identification[C]//In Proceedings of the IEEE International Conference on Computer Vision (ICCV). 2019: 8321-8330.

[18] VAN DEN OORD A, LI Y, VINYALS O. Representation learning with contrastive predictive coding[J/OL]. CoRR, 2018, abs/1807.03748. http://arxiv.org/abs/1807.03748.

[19] GE Y, ZHU F, CHEN D, et al. Self-paced contrastive learning with hybrid memory for domain adaptive object re-id[J]. Advances in Neural Information Processing Systems (NeurIPS), 2020, 33: 11309-11321.

[20] DAI Z, WANG G, ZHU S, et al. Cluster contrast for unsupervised person re-identification. arxiv 2021[J]. arXiv preprint arXiv:2103.11568, 2021.

[21] CHEN H, LAGADEC B, BREMOND F. Ice: Inter-instance contrastive encoding for unsupervised person re-identification[C]//In Proceedings of the IEEE International Conference on Computer Vision (ICCV). 2021: 14960-14969.

[22] LIU Y, ZHOU W, XI M, et al. Vision meets wireless positioning: Effective person re-identification with recurrent context propagation[C]//In Proceedings of the ACM International Conference on Multimedia (ACM MM). 2020: 1103-1111.

[23] LIU Y, ZHOU W, XIE Q, et al. Unsupervised person re-identification with wireless positioning under weak scene labeling[J/OL]. CoRR, 2021, abs/2110.15610. https://arxiv.org/abs/2110.15610.

[24] FELZENSZWALB P F, GIRSHICK R B, MCALLESTER D, et al. Object detection with discriminatively trained part-based models[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 2010, 32(9): 1627-1645.

[25] 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.

[26] ZHENG L, SHEN L, TIAN L, et al. Scalable person re-identification: A benchmark[C]//In Proceedings of the IEEE International Conference on Computer Vision (ICCV). 2015: 1116-1124.

[27] ZHENG Z, ZHENG L, YANG Y. Unlabeled samples generated by gan improve the person re-identification baseline in vitro[C]//In Proceedings of the IEEE International Conference on Computer Vision (ICCV). 2017: 3754-3762.

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

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

[30] REN S, HE K, GIRSHICK R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks[C]//In Advances in Neural Information Processing Systems (NeurIPS). 2015: 91-99.

[31] WU A, ZHENG W, GONG S, et al. Rgb-ir person re-identification by cross-

modality similarity preservation[J]. International Journal of Computer Vision(IJCV), 2020, 128: 1765-1785.

[32] XIE Q, ZHOU W, QI G J, et al. Progressive unsupervised person re-identification by tracklet association with spatio-temporal regularization[J]. IEEE Transactions on Multimedia (TMM), 2020.

[33] REDMON J, FARHADI A. Yolov3: An incremental improvement[J/OL]. CoRR, 2018, abs/1804.02767. http://arxiv.org/abs/1804.02767.

[34] LI B, YAN J, WU W, et al. High performance visual tracking with siamese region proposal network[C]//In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2018: 8971-8980.

[35] KINGMA D P, BA J. Adam: A method for stochastic optimization[C]//In Proceedings of the International Conference on Learning Representations (ICLR). 2015.

[36] ESTER M, KRIEGEL H P, SANDER J, et al. A density-based algorithm for discovering clusters in large spatial databases with noise[C]//KDD'96: In Proceedings of the Second International Conference on Knowledge Discovery and Data Mining. Portland, Oregon: AAAI Press, 1996: 226–231.

[37] ROSVALL M, AXELSSON D, BERGSTROM C T. The map equation[J]. The European Physical Journal Special Topics, 2009, 178: 13-23.

[38] XUAN S, ZHANG S. Intra-inter camera similarity for unsupervised person re-identification[C]//In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2021: 11926-11935.

[39] YANG F, ZHONG Z, LUO Z, et al. Joint noise-tolerant learning and meta camera shift adaptation for unsupervised person re-identification[C]//In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2021: 4855-4864.