



浙江工业大学

硕士学位论文

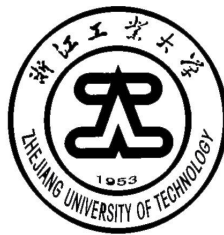
基于深度学习的跨模态行人重识别研究

作者姓名	林建武
指导教师	赵云波 教授
学科专业	控制科学与工程
学位类型	工学硕士
培养类别	全日制学术型硕士
所在学院	信息工程学院

提交日期：2020 年 06 月

Research on Cross-modality Person Re-identification Based on Deep Learning

Dissertation Submitted to
Zhejiang University of Technology
in partial fulfillment of the requirement
for the degree of
Master of Engineering



by
Jian-wu LIN

Dissertation Supervisor: Prof. Yun-bo ZHAO

Jun., 2020

浙江工业大学学位论文原创性声明

本人郑重声明：所提交的学位论文是本人在导师的指导下，独立进行研究工作所取得的研究成果。除文中已经加以标注引用的内容外，本论文不包含其他个人或集体已经发表或撰写过的研究成果，也不含为获得浙江工业大学或其它教育机构的学位证书而使用过的材料。对本文的研究作出重要贡献的个人和集体，均已在文中以明确方式标明。本人承担本声明的法律责任。

作者签名：林建武

日期：2020年5月

学位论文版权使用授权书

本学位论文作者完全了解学校有关保留、使用学位论文的规定，同意学校保留并向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅和借阅。本人授权浙江工业大学可以将本学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存和汇编本学位论文。

本学位论文属于 1、保密□，在一年解密后适用本授权书。

2、保密□，在二年解密后适用本授权书。

3、保密□，在三年解密后适用本授权书。

4、不保密☒。

(请在以上相应方框内打“√”)

作者签名：林建武

日期：2020年5月

导师签名：赵海

日期：2020年5月

中图分类号 TH741

学校代码 10337

UDC 681.2

密级 公开

研究生类别 全日制专业型硕士研究生



浙江工业大学

工学硕士学位论文

基于深度学习的跨模态行人重识别研究

Research on Cross-modality Person Re-identification Based on
Deep Learning

作者姓名 林建武

第一导师 赵云波 教授

学位类型 工学硕士

学科专业 控制科学与工程

培养单位 信息工程学院

研究方向 计算机视觉

答辩日期: 2020 年 5 月 17 日

基于深度学习的跨模态行人重识别研究

摘 要

行人重识别在安防领域有重要的应用，而跨模态行人重识别可以在异质数据中匹配相同身份的行人，可以解决全天候监控的问题。相比于手工设计的特征匹配方法，基于深度学习的跨模态行人重识别算法在精度、鲁棒性方面都有很大的提升，因此基于深度学习的跨模态行人重识别算法在学术界和产业界引起了广泛关注。

本文针对跨模态行人重识别任务，从神经网络结构、损失函数两个方面对基于深度学习的跨模态行人重识别算法进行分析与设计。与现有论文相比，本文设计的方法在跨模态行人重识别的两个公开数据集 SYSU-MM01 和 RegDB 中取得了领先的精度。本文的主要工作包括以下三个方面：

(1) 设计了一种基于困难五元组的特征提取框架，该框架具有高精度、可扩展性的特点。该框架使用改进的单模态行人重识别网络作为特征提取模块，现有的单模态行人重识别模型相比于分类网络有更强的特征提取能力，同时具有高度可扩展性。此外，本文设计了困难五元组损失函数以及相应的困难五元组采样方法。困难五元组损失函数由全局三元组和跨模态三元组损失函数组成，前者同时处理了模态内变化和模态间变化，由于模态间变化更大，后者增强前者对模态间变化的关注力度，该损失函数在训练效率、精度等方面有显著提高。在该框架中，我们级联困难五元组损失和身份损失函数，进一步提升了精度。

(2) 设计了一种基于多粒度特征的双通道网络。两个通道由两个结构相同但参数不共享的特征提取模块构成，两个特征提取模块分别提取红外图像与可见光图像的行人特征，每个特征提取模块由一个全局特征分支和两个局部特征分支组成，融合全局特征和局部特征的策略使网络具有提取多粒度特征的能力，使模型可以关注到行人不同部位的细节特征。最后通过身份损失和异质中心损失函数来提升特征向量的判别性。

(3) 设计了一种基于梯度协调机制的身份损失函数，一定程度上缓解了易样本和由于模态差异过大造成的离群点对模型训练的影响。在模型结构不变的情况下，通过梯度协调机制辅助身份损失来提升训练结果，该方法具有扩展性。

关键词： 跨模态行人重识别，跨模态图像检索，智能安防，度量学习

RESEARCH ON CROSS-MODALITY PERSON RE-IDENTIFICATION BASED ON DEEP LEARNING

ABSTRACT

Person re-identification has important applications in the field of security, and cross-modality person re-identification matches pedestrians with the same identity in heterogeneous data to solve the problem of 24-hours monitoring. Compared to hand-designed feature, the deep learning-based cross-modality person re-identification algorithm has greatly improved in terms of accuracy and robustness, so it has important significance for cross-modality person re-identification.

This paper aims at cross-modality person re-identification tasks, analyzes and designs cross-modality person re-identification algorithms based on deep learning from the aspects of neural network structure and loss function. Compared with existing papers, the method designed in this paper has achieved leading accuracy in two public datasets (SYSU-MM01 and RegDB) for cross-modality person re-identification. The main work of this article includes the following three aspects:

(1) A feature extraction framework based on hard pentaplet was designed. The framework has the characteristics of high accuracy and high scalability. The framework uses the improved single-modality person re-identification network as feature extraction module. Because the existing single-modality person re-identification model has stronger feature extraction capability than classification network, it also has high scalability. In addition, this paper designs hard pentaplet loss and a corresponding hard pentaplet sampling method. The hard pentaplet loss consists of a hard global triplet and a hard cross-modality triplet loss. The former handles both intra-modality and inter-modality variations. Since the inter-modality variations are more serious, the latter enhances the former's focus on inter-modality variations, and the loss function significantly improves training efficiency and accuracy. In this framework, we cascade the hard pentaplet loss and identity loss functions, further improving the accuracy.

(2) A dual-channel network based on multi-granularity features was designed. The two channels are composed of two feature extraction modules, which has the same structure and different parameters. The two feature extraction modules extract the pedestrian features of the infrared images and the visible images. Each feature extraction module consists of a global feature branch and two local feature branches.

The strategy of composition and fusion of global features and local features makes the network have the ability to extract multi-granularity features, so that the model can pay attention to the detailed features of different parts of the pedestrian. Finally, the discriminability of feature vectors is improved through identity loss and heterogeneous center loss functions.

(3) The role of the gradient harmonizing mechanism in the field of cross-modality person re-identification was studied, and the effects of easy samples and outliers caused by excessive modal differences on model training are addressed. The identity loss based on the gradient coordination mechanism can improve the training results while keeping the model structure unchanged. This method is scalable and universal.

KEY WORDS: Cross-modality person re-identification, cross-modality image retrieval, intelligent security, metric learning

目 录

摘 要.....	I
ABSTRACT.....	II
插图清单.....	VII
表格清单.....	VIII
第一章 绪 论.....	1
1.1 研究背景和意义.....	1
1.2 国内外研究现状.....	2
1.2.1 单模态行人重识别研究现状.....	3
1.2.2 跨模态行人重识别研究现状.....	4
1.3 实验数据集和评测指标.....	6
1.3.1 跨模态行人重识别数据集.....	6
1.3.2 评测指标.....	7
1.4 本文组织架构.....	9
1.5 本章小结.....	9
第二章 相关技术介绍.....	11
2.1 人工神经网络.....	11
2.1.1 M-P 神经元模型.....	11
2.1.2 多层感知机.....	13
2.1.3 反向传播算法.....	14
2.2 卷积神经函数.....	15
2.2.1 卷积层.....	15
2.2.2 激活层.....	16
2.2.3 池化层.....	17
2.2.4 全连接层.....	18
2.3 行人重识别方法.....	19
2.3.1 基于全局特征的 Re-ID 方法.....	19

2.3.2 基于局部特征的 Re-ID 方法.....	19
2.3.3 基于视频序列的 Re-ID 方法.....	19
2.4 本章小结.....	19
第三章 基于困难五元组的特征提取框架的设计.....	21
3.1 特征提取框架的设计.....	21
3.1.1 改造单模态行人重识别网络.....	22
3.1.2 困难五元组损失函数.....	23
3.1.3 级联身份损失.....	26
3.2 实验与分析.....	26
3.2.1 实验设置与环境.....	26
3.2.2 实验对比.....	27
3.2.3 分析级联损失函数的作用.....	28
3.2.4 卷积特征可视化.....	31
3.3 本章小结.....	32
第四章 基于多粒度特征的双通道网络的设计.....	33
4.1 双通道网络结构设计.....	33
4.1.1 多粒度特征模块.....	34
4.1.2 损失函数模块.....	35
4.2 实验与分析.....	36
4.2.1 实验设置.....	36
4.2.2 实验结果.....	36
4.2.3 排序结果可视化.....	37
4.3 本章小结.....	38
第五章 基于梯度协调机制的身份损失函数的设计.....	39
5.1 引言.....	39
5.2 梯度协调机制.....	40
5.3 实验与分析.....	42
5.3.1 梯度模长可视化.....	42
5.3.2 实验对比.....	43
5.4 本章小结.....	44

第六章 总结与展望.....45

 6.1 总结.....45

 6.2 展望.....46

参考文献.....47

致 谢.....50

作者简介.....51

 1 作者简历.....51

 2 攻读硕士学位期间发表的学术论文.....51

 3 参与的科研项目及获奖情况.....51

 4 发明专利.....51

学位论文数据集.....53

插图清单

图 1-1 行人重识别顶会论文数量统计.....	2
图 1-2 典型的跨模态行人重识别网络结构.....	5
图 1-3 SYSU-MM01 数据集.....	6
图 1-4 RegDB 数据集.....	7
图 1-5 不同模型的检索结果.....	8
图 2-1 M-P 神经元模型.....	11
图 2-2 典型的激活函数.....	12
图 2-3 单隐层感知机.....	13
图 2-4 无填充的卷积过程.....	16
图 2-6 ReLU 和 PReLU 激活函数.....	17
图 2-7 两种典型的池化方式.....	18
图 2-8 全连接层.....	18
图 3-1 HPILN 特征提取框架.....	22
图 3-2 典型的单模态行人重识别网络.....	23
图 3-3 三元组损失函数.....	24
图 3-4 困难五元组损失函数.....	26
图 3-5 损失函数的对比结果.....	29
图 3-6 不同损失函数的 CMC 曲线.....	30
图 3-7 损失函数特征可视化.....	31
图 3-8 卷积特征激活图.....	31
图 4-1 基于多粒度特征的双通道网络结构图.....	34
图 4-2 排序可视化.....	38
图 5-1 离群点特征分布示意图.....	39
图 5-2 离群点特征分布真实图.....	40
图 5-3 不同梯度模长的样本分布情况.....	43
图 5-4 不同训练阶段的 Rank1 变化图.....	44

表格清单

表 3-1 图片输入尺寸和特征维度.....	27
表 3-2 HPILN 在 ASSS 模式下的实验结果.....	28
表 3-3 HPILN 框架在 ASMS、ISSS、ISMS 模式下的实验结果.....	28
表 4-1 D-MGN 网络在 SYSU-MM01 数据集的实验结果.....	37
表 4-2 D-MGN 网络在 RegDB 数据集的实验结果.....	37
表 5-1 梯度协调机制在 SYSU-MM01 上的实验结果.....	43

第一章 绪 论

1.1 研究背景和意义

随着老人走失、儿童拐卖、犯罪事件的频发，人工寻找目标人物的效率已经无法满足人们的需求。便捷的交通使目标行人在短时间内能够快速移动，复杂的城市建筑遮挡了人们的视野，这些因素加大了人工寻找目标的难度。近年来，随着平安城市、天网工程、雪亮工程的推进实施，中国城市和农村的监控设备数目增长迅速。据统计，截止 2019 年中国有两亿部监控摄像头正投入使用，预计 2020 年将达到 6 亿部。人工在监控视频中寻找目标人物是十分耗时耗力的，比如在轰动一时的周克华案件中，长沙市 1000 名民警耗费两个月时间查看 30 万 G 的监控视频，最终在海量视频中找到了嫌疑人。如今，随着监控设备的广泛分布以及深度学习算法的发展，使得利用图像算法在监控画面中寻找目标人物成为可能^[1]。

人脸识别是根据图像辨别身份的常用方法，但应用于监控场景中存在难以避免的缺陷。人脸识别按照使用方式划分为主动式人脸识别和被动式人脸识别，前者表示测试者主动配合机器进行人脸识别，如闸机的人脸验证，后者表示测试者在不知情的情况下被机器识别，如在监控画面中寻找目标人物。然而，被动式人脸识别应用于寻找目标人物存在两方面缺陷。首先，行人不会主动配合监控摄像头的图像采集，面部容易受到遮挡，又或者行人背对着监控摄像头，导致监控摄像头常常无法拍摄到面部；其次，由于人脸面积在监控画面中占比较小，再加上监控探头的分辨率有限，导致监控摄像头拍摄到的人脸不够清晰，从而严重影响人脸识别算法的准确度。因此，人脸识别在寻找目标人物的任务中存在非常大的限制。

考虑到人脸识别的局限性，研究人员提出根据行人的全身图像来进行身份识别，这种技术称为行人重识别（Person Re-identification, Re-ID）。行人重识别也称行人再识别，是利用计算机视觉判断某一行人是否在监控摄像头网络下出现过的技术^[2]。相比与人脸识别，行人重识别可利用的图像更大、可提取的信息更多。行人重识别可以弥补目前固定的摄像头的视觉局限，在跨摄像头中搜索目标行人，广泛应用于智能视频监控、智能安保等领域。由于不同摄像设备之间的差异，行人外观易受穿着、尺度、遮挡、姿态和视角等影响，使得行人重识别成为计算机视觉领域中一个既具有研究价值同时又极具挑战性的热门课题。

在不同场景需要使用不同光谱类型的监控摄像头，常见的摄像头有可见光摄像头、红外摄像头、热成像摄像头。可见光摄像头在黑暗环境中不能拍摄到清晰的图像，而红外摄像头能在黑暗环境中正常工作。热成像摄像头通过捕捉物体散发的热量来成像，常常被用于夜视场景。不同光谱摄像头拍摄到的图像属于不同模态，跨模态行人重识别是判断某一行人是否在跨模态摄像头中出现过的技术。可见光-红外跨模态组合、可见光-热成像跨模态组合是实际应用中最常见的两种组合，也是当前跨模态行人重识别研究的热点。研究跨模态行人重识别的目的是使得不同模态的行人图像能够相互匹配，充分利用多种摄像头拍摄到的监控数据，达到 24 小时监控的目标。单模态行人重识别无法匹配异质数据中的行人，跨模态行人重识别解决了全天候的监控问题，在安防监控、智能视频监控领域有重要的应用价值。

1.2 国内外研究现状

由于单模态行人重识别的研究起步较早，且单模态与跨模态行人重识别的目标都是通过图像或视频寻找目标人物，因此跨模态行人重识别的方法往往是在单模态行人重识别方法上进行改进和创新。在 2006 年的计算机视觉顶会 CVPR 中，行人重识别被正式提出，由于该任务的重要性和挑战性，吸引了大批学者进行研究。我们统计了行人重识别十年间在计算机视觉顶级会议（CVPR、ICCV 和 ECCV）的收录情况，如图 1-1 所示，行人重识别在顶级会议中的录用数量逐年上升，这意味着行人重识别领域越来越受研究人员重视。

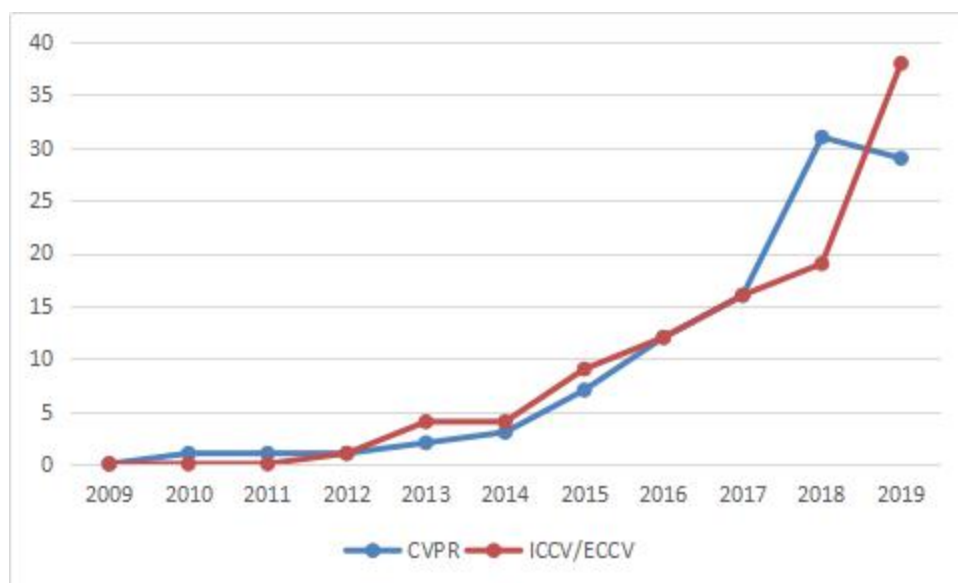


图 1-1 行人重识别顶会论文数量统计

Figure 1-1. Number of person re-identification papers

起初, 由于早期行人重识别数据集^[3-5]只采集了可见光模态的图像, 因此之前的研究成果主要集中在单模态行人重识别领域。直到 2017 年, 中山大学的学者在 ICCV (计算机视觉顶会) 上发表了第一篇跨模态行人重识别论文^[6], 并且公开了首个跨模态行人重识别数据集 (SYSU-MM01)。此后, 跨模态行人重识别成为了行人重识别领域中的热点方向。

1.2.1 单模态行人重识别研究现状

行人重识别的难点在于如何将图像或视频转化为在度量空间中具有判别性的特征向量。起初, 手工设计的特征描述子是单模态行人重识别主流的特征表示方法, 通过手工设计的特征描述子提取行人特征, 然后用度量公式来计算两张图片的相似性。在深度学习出现之后, 研究人员通过设计卷积神经网络来自动提取特征, 并设计合理的损失函数来优化特征向量。下文中我们将介绍单模态行人重识别中手工特征描述子、深度学习两方面的发展。

在研究行人重识别的早期阶段, 手工设计的特征描述子是刻画行人图像共性特征的主流方法。常使用颜色和纹理信息来设计特征描述子, 其中基于颜色信息的特征描述子占主要部分。基于颜色信息设计的特征描述子主要有 RGB、HSV 和 LAB 色彩空间^[7, 8]的颜色直方图, 而基于纹理信息设计的特征描述子主要有 SIFT^[9]和 RHSP^[10]。RGB、HSV^[11]和 LAB 是三种不同的色彩空间, 将图像转化到某种色彩空间后, 以颜色直方图的形式来统计不同像素值的分布。颜色直方图可以较好地处理行人姿态的变化, 进一步的, 加权颜色直方图对光照变化有更强的抗干扰性。在纹理特征方面, SIFT (Scale-Invariant Feature Transform) 在空间尺度计算极值点, 平滑区域变化不大, 而纹理区域存在大量极值点, SIFT 根据极值点筛选出关键点, 统计每个关键点邻域范围内的梯度分布, 并以特征向量来表示。SIFT 的优点是具有旋转不变性, 对图像尺度变化有鲁棒性, 并且由于计算像素梯度的缘故, 对光线变化的容忍度也很高。后来研究人员将颜色特征与纹理特征融和, 得到更鲁棒的特征。比如, Farenzena^[10]等人将行人身体从背景中分割出来, 并计算各个身体部位的加权颜色直方图、重复结构区域 (Recurrent High-Structured Patches, RHSP) 作为行人特征。使用加权颜色直方图的好处在于能够处理行人的姿势变化和光照强度不同导致的明暗变化, 而重复结构区域是一种纹理特征, 具有姿态和旋转不变性, 融合两种特征提升了识别效果。

深度学习是一种从数据中自动提取特征的算法, 2012 年深度学习算法在 ImageNet 图像分类竞赛上大放异彩^[12], 基于深度学习的行人重识别算法成为新的发展方向。基于深度学习的行人重识别算法的重点在于如何设计卷积神经网络结构^[13]和损失函数, 卷积神经网络是深度学习中自动提取图像特征的一种结构, 其组成结构、深度都会影响特征提取能力。在训练阶段, 损失函数决定了模型中参数的调整方向, 因此合理的损失函数有助于卷积神经网络更好地提取

行人特征。2014 年，两篇基于深度学习的行人重识别论文首次被发表，Yi 等人^[14]和 Li 等人^[15]都提出使用孪生卷积神经网络结构来计算两张行人图片的相似度。Yi 等人^[14]提出的孪生网络输入两张图片，前半部分通过两个结构对称的子分支网络提取两个行人的特征向量，后半部分使用余弦距离计算两个特征向量的相似度，作者实验了均方损失、指数损失、二项偏差损失等损失函数在该网络中的效果。此外，上述的子网络针对行人特点进行了设计，将行人水平分割成上、中、下三个部位，用卷积神经网络对每个部位分别进行特征提取。这种水平分割的方案被后来的研究人员广为使用，Sun 等人^[16]提出的 PCB（Part-based Convolutional Baseline）将特征图分割为 6 个水平条，并使用损失函数对每个水平条进行监督，基于水平分割的局部特征能很好地关注不同部位的细节特征。Wang 等人^[17]提出的多粒度网络（Multiple Granularity Network, MGN）使用三个通道的形式将全局特征和局部特征融合在一起，第一个网络通道提取全局特征，第二个通道将特征图水平分割成上、下两个子块，第三个通道将特征图水平分割成上、中、下三个子块，先后使用全局最大池化和 1*1 卷积操作对每个子块进行降维操作，并级联降维后的特征作为融合特征。虽然 PCB 和 MGN 沿用了文献^[14]中水平分块的思想，但是却没有采用孪生网络的结构来获取两张图片的相似度，而是输入一张图片，输出为该图片的特征向量，图像的相似度用特征向量的欧式距离来表示，这种方式称为表征学习。原因在于孪生网络在现实应用中存在计算冗余的问题，在实际应用中，往往将库中的行人图片转化为特征向量进行保存，表征学习的做法避免了对库中行人反复进行特征提取的工作，从而提高效率。在损失函数方面，行人重识别的损失函数大量地借鉴了人脸识别中的损失函数，主要有身份损失（Identity loss）^[18]、验证损失（Verification loss）^[19]、对比损失（Contrastive loss）^[20]、三元组损失（Triplet loss）^[21]，以及困难三元组损失（Hard triplet loss）^[22]，其中身份损失和困难三元组损失使用最为广泛。

1.2.2 跨模态行人重识别研究现状

跨模态行人重识别是判断某一行人是否在跨模态摄像头中出现过的技术^[23]，现有的跨模态行人重识别研究主要集中在可见光-红外行人重识别（RGB-IR Re-ID）、可见光-热成像行人重识别（VT Re-ID）。红外摄像头和热成像摄像头都能够在夜间工作，它们的区别在于前者主动发射并收集红外光以获取图像，而后者通过捕捉人体散发的热量来获取图像。随着红外摄像头和热成像摄像头的普及，研究人员意识到单模态行人重识别的局限性。2017 年，Wu 等人^[6]公开了首个 RGB-IR 跨模态行人重识别数据集 SYSU-MM01，并为该数据集设计了一套评测标准。Wu 等人的文章中介绍了三种跨模态图像识别领域中常用的神经网络结构：单通道结构、双通道结构、非对称全连接层结构，如图 1-2 所示：

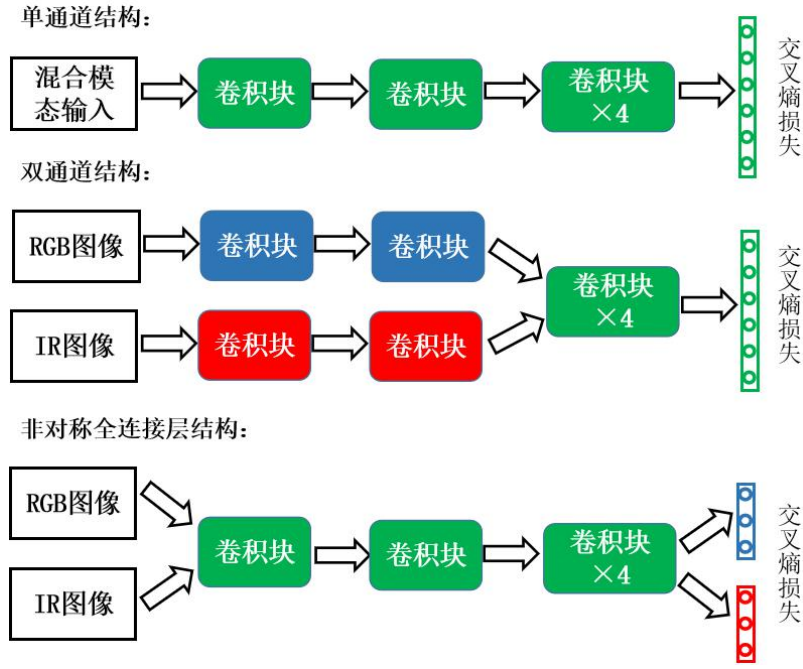


图 1-2 典型的跨模态行人重识别网络结构

Figure 1-2. Typical cross-modality Re-ID network structure

单通道结构中所有权值都是共享的，不区分模态差异。双通道结构广泛应用于跨模态图像检索领域^[24]，它的前半部分权值不共享，使用两个通道来提取不同模态的行人特征，后半部分为共享权值的卷积块，用来提取有判别性的行人特征。非对称全连接层结构的思想与双通道类似，只是将权值不共享的部分设置在最后一层的全连接层上。

此后，大量的跨模态行人重识别方法被提出，其中基于双通道结构的算法^[25-30]占多数。在双通道结构的算法中，除了卷积层的组合有差异以外，损失函数的差别最大。文献^[25]提出双向限制的排序损失函数，该损失函数基于三元组损失^[21]针对跨模态行人重识别任务进行改进，该损失函数同时考虑了模态间的变化和模态内的变化。Ye 等人^[30]将对比损失引入到跨模态行人重识别中，并融合身份损失以提高精度，此外，作者将 Nguyen 等人^[31]公开的跨模态行人识别数据集（DBPerson-Recog-DB1）进行改造，制定出首个可见光-热成像跨模态行人重识别数据集 RegDB 的实验标准和验证规则。Zhu 等人^[29]提出的异质中心损失函数和 Ye 等人^[26]提出的双向中心排序损失函数都改造自人脸识别中的中心损失函数^[32]，在训练过程中计算每一个类的类中心，样本在特征空间中与类中心的欧式距离作为中心损失值，中心损失有助于增加类内聚合度和类间的分离度。

生成对抗网络（Generative Adversarial Network, GAN）^[33]在跨模态行人重识别领域中效果显著。Dai 等人^[34]提出一种基于对抗学习的端对端训练方法，使用判别器来判断图像的模态，并使用三元组损失函数和身份损失函数来训练

生成器，经过对抗训练，不同模态的行人特征被映射到同一特征空间。Wang 等人^[35]通过 CycleGAN^[36] 将可见光图像生成相同身份的红外图像，在特征对齐模块中将生成的红外图像和真实的红外图像通过同一个卷积神经网络分别映射成两个特征图，在对抗学习过程中，判别器负责判断红外图像是否为生成的，使得生成器生成的红外图像更具真实感。文献^[37]提出将可见光图像和红外图像互相转换，通过变分自编码器将两种模态的图像转为中间隐向量，再用解码器生成另一种模态的图像。之后，将真实的可见光/红外图像与生成的红外/可见光模态图像在通道层级上融合为一张四通道的图（可见光图像有三个通道，红外光只有一个通道），再将融合后的图使用身份损失和三元组损失函数进行训练，得到有判别性的特征向量。

1.3 实验数据集和评测指标

1.3.1 跨模态行人重识别数据集

目前，跨模态行人重识别数据集有两个：SYSU-MM01^[6]和 RegDB^[30]，我们在下文中的实验中将使用这两个数据集来验证算法的性能。

SYSU-MM01 是首个公开的可见光-红外跨模态行人重识别（RGB-IR Re-ID）数据集。该数据集包含 491 个行人的 287628 张可见光图像和 15792 张红外图像，这些图像由四个可见光相机（相机 1，2，4，5）和两个红外相机拍摄（相机 3，6），采集的图像如图 1-3 所示。可见光相机安装在明亮处，而红外相机则安装在昏暗的地方，以保证数据集采集的图像与实际情况相符合。为了防止室内、室外光线不同造成影响，相机 1、2、3 被安装在室内，相机 4、5、6 被安装在室外。SYSU-MM01 数据集包含训练集和测试集两个子集，训练集中包含 395 个行人，测试集中包含 96 个行人。



图 1-3 SYSU-MM01 数据集

Figure 1-3. The SYSU-MM01 dataset

直观地看，SYSU-MM01 数据集中的可见光图像和红外图像差异很大，即使是人眼观察，也很难分辨其身份。“person1”穿着蓝色格子衬衫，然而在红外相机 3、6 中，格子纹理消失，颜色由蓝变灰，并且由于相机视角不同和人体姿态变化的因素，加大了识别的难度。“person2”和“person3”是两个穿着红色上衣的男子，在可见光相机和红外相机中都非常难以区分，较小的类间模态内差异也增大了识别难度。“person4”和“person5”分别为穿着白色和黑色裙子的女子，他们在可见光相机中很容易被区分开来，但是在红外相机 3、6 中非常难以区分。毫无疑问，SYSU-MM01 是一个极具挑战的数据集。

RegDB 数据集采集自双相机装置，即该装置集成了可见光相机和热成像相机。该数据集包含 412 个行人，每个行人采集 10 张可见光图像和 10 张热成像图像。由于相机系统中可见光相机和热成像相机的位置非常相近，因此采集的可见光图像和热成像图像中的行人姿势、角度、轮廓非常相近。RegDB 采集的图像如图 1-4 所示。热成像相机拍摄的图像信息缺失非常严重，只保存了轮廓信息，并且由于外界发热物的干扰，破坏热成像图像的行人轮廓，因此可见光-热成像跨模态行人重识别的 RegDB 数据集极具挑战性。



图 1-4 RegDB 数据集
Figure 1-4. The RegDB dataset

1.3.2 评测指标

在介绍上述两个数据集的评测标准之前，首先介绍行人重识别领域中两个重要的评测指标：CMC 曲线和 mAP。

CMC 曲线全称为累计匹配曲线（Cumulative Match Characteristic Curve），是图像检索领域的重要指标。行人重识别在测试时，会给定一个查询目标库（query）和候选行人库（gallery），分别计算查询目标库中每一个人与候选行人库中所有人的相似度，并将相似度由大到小排序，相似度的排列结果用一个二维矩阵表示。CMC 曲线的横坐标为 n ，纵坐标是 $Rank_n$ 的值， Top_n 表示相似度向量前 n 项中有与查询目标相同身份的概率（0 或者 1），由于测试集 query 中有 P 张查询图片， $Rank_n$ 为这 P 张图片的 $Top-n$ 的平均值。

$$Rank_n = \frac{1}{P} \sum_{m=1}^P Top_m \quad (1-1)$$

当 gallery 库中有多张相同身份的图片时，Rank_n 的值不能很好的体现模型的好坏，Zheng 等人^[3]提出使用平均检索精度（mean average precision, mAP）进行评估。图 1-5 表示两个模型对图片 A 的检索结果，两个模型都认为 A1 与 A 是最相似的，因此两个模型的 Top_n 相等（都为 1），然而，模型 1 显然要比模型 2 的性能要好，因为在模型 1 的排序结果中，A2 的名次更靠前，因此需要 mAP 来衡量模型的好坏。AP 的计算公式为：

$$AP = \frac{\left(\frac{p_1}{1} + \frac{p_2}{2} + \dots + \frac{p_i}{i} + \dots + \frac{p_n}{n}\right)}{k} \quad (1-2)$$

其中， n 表示候选行人库（gallery）中行人的数量， k 表示候选行人库中与查询目标身份相同的图片数量， p_i 表示排序结果中与目标行人匹配的次序数。图 1-5 中，模型 1 的 AP 等于 $(1/1+2/2)/2=1$ ，模型 2 的 AP 等于 $(1/1+0/2+0/3+0/4+0/5+2/6)/2=0.67$ 。

mAP 为查询目标库中所有图片的检索精度（AP）的平均值：

$$mAP = \frac{1}{y} \sum_{j=1}^y AP_j \quad (1-3)$$

其中， y 表示查询目标库中图片的数量。

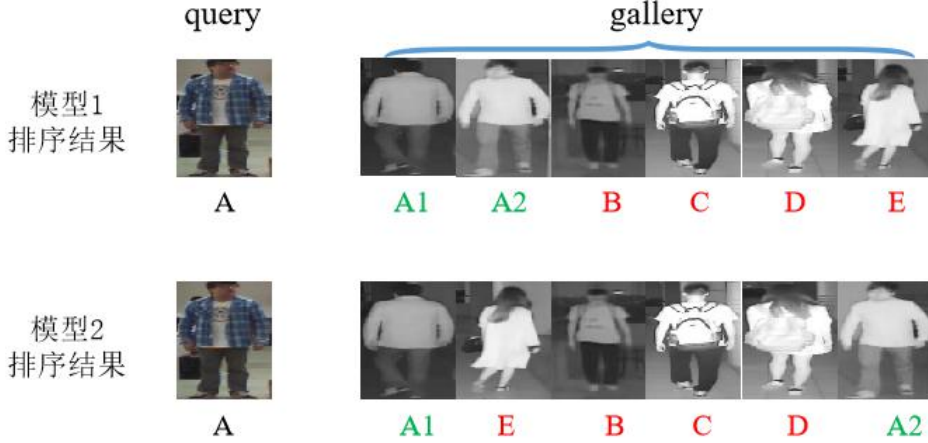


图 1-5 不同模型的检索结果

Figure 1-5. Search results for different models

在 SYSU-MM01 和 RegDB 数据集中，一个人在候选行人库中具有多个图像，这意味着一个查询图像可能具有多个匹配目标。因此，除了 CMC 曲线外，还采用平均检索精度（mAP）作为评估指标，但是它们的评测方法不同。

SYSU-MM01 数据集的评测方法如下：在训练阶段，所有的训练集图片可以用于训练。在测试阶段，可见光图片作为候选行人库（gallery），红外图片作为查询目标库（query）。根据摄像头的安装位置，可分为两种搜索模式：全

全局搜索模型（all-search）和室内搜索模式（indoor-search）。全局搜索模式中，测试集包含了室内和室外的所有图片，而在室内搜索模式中，测试集仅包含室内拍摄的图片。每一种搜索模式中又包含多发命中（multi-shot）和单一命中（single-shot）两种设置，对于候选行人库中的每个身份，我们随机选取 1 张或 10 张可见光图片作为单一命中和多发命中两种设置的候选行人库。无论哪种设置，查询目标库中所有的红外图片被选取测试。

RegDB 数据集的评测方法如下：将数据集随机分为两半，一半用于训练，另一半用于测试。在测试阶段，将来自某个模态的图像用作候选行人库（gallery），将另一个模态的图像用作查询目标库（query）。重复上述步骤进行 10 次试验，最终取十次实验的平均值，使结果更具可信性。

1.4 本文组织架构

本文是基于深度学习的跨模态行人重识别研究，重点对单通道网络结构、双通道网络结构、跨模态行人重识别损失函数、平衡难易样本等方面进行研究。本文分六个章节进行阐述，其文章组织架构如下：

第一章：绪论。介绍了课题的研究背景与意义、国内外研究现状、实验数据集和评测指标。

第二章：相关技术介绍。本章介绍了本文用到的一些常用方法和公式。

第三章：基于困难五元组的特征提取框架的设计。本章提出了一种特征提取框架，通过改造单模态行人重识别网络用于跨模态行人重识别中，提出一种困难五元组损失函数，并级联身份损失函数。实验证明该方法在发表时取得最高精度。

第四章：基于多粒度特征的双通道网络的设计。本章设计了一种双通道网络结构，每个通道使用多粒度分割的方案提取全局特征和局部特征，使用身份损失和异质中心损失函数进行训练，通过实验证明了该方法的可靠性。

第五章：基于梯度协调机制的身份损失函数的设计。本章使用梯度协调机制改进身份损失函数，以减少训练过程中易样本和离群点的损失值，加深模型的收敛程度。

第六章：结论与展望。总结全文工作，分析先前工作的优势与不足，并展望未来可发展的方向。

1.5 本章小结

本章首先介绍了跨模态行人重识别的研究背景和意义，从单模态行人重识别和跨模态行人重识别两个方面介绍了近年来国内外的研究现状。针对跨模态

行人重识别中的两个数据集，我们介绍了本文实验中使用的数据集的构成、评测指标和评测方法。最后介绍了本文章节的组织分布。

第二章 相关技术介绍

本文基于深度学习对跨模态行人重识别进行研究，不仅需要对深度学习中卷积神经网络进行研究，而且还需要了解行人重识别中常用的损失函数。本章介绍了早期人工神经网络的神经元模型和多层感知机，以及神经网络在训练过程中使用的反向传播算法。同时也介绍了当前流行的卷积神经网络中的卷积层、激活层、池化层和全连接层等结构。通过阅读本章可以更好地理解后文中的内容。

2.1 人工神经网络

人工神经网络是按照一定规则将神经元组合而成的网络结构，是一种信息处理的数学模型。人工神经网络仿照生物神经网络结构，神经元模型也仿照生物神经元的构造。生物神经网络由多个神经元连接而成，当生物感受器受到外界刺激时，相应的神经元以电位差的形式产生“兴奋”状态，当电位差超过一定阈值时，这种“兴奋”状态将在神经元间传递。

本节首先介绍 M-P 神经元模型、多层感知机，再演算反向传播（BP）算法的推导过程。

2.1.1 M-P 神经元模型

1943 年，McCulloch 和 Pitts^[38]将生物神经元抽象成逻辑运算模型：M-P 神经元模型，如图 2-1 所示。

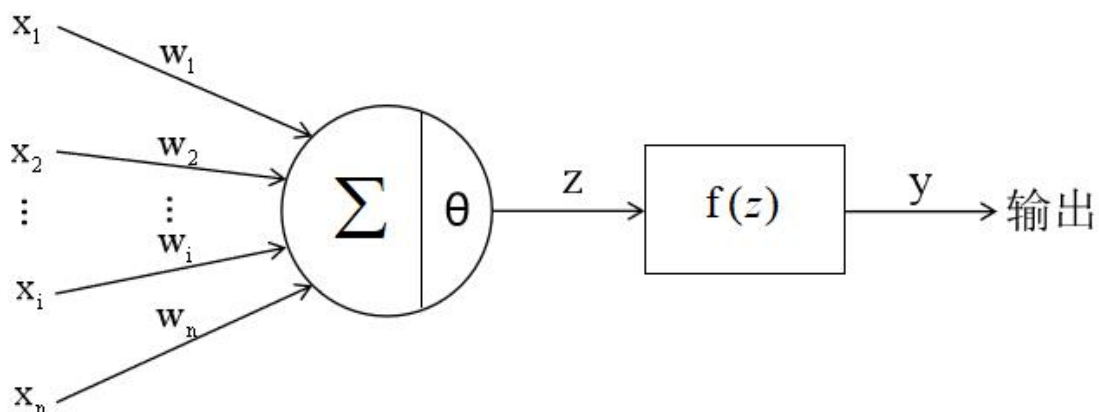


图 2-1 M-P 神经元模型

Figure 2-1. M-P neuron model

M-P 神经元模型用公式描述为：

$$y = f(z) = f\left(\sum_{i=1}^n w_i x_i - \theta\right) \quad (2-1)$$

其中, x_i 表示该模型从其他神经元传递来的输入信号, 权值和阈值分别用 w_i 和 θ 表示。仿照生物神经元的运行模式, 对加权后的输入信号进行求和, 并减去阈值 θ , 最后通过激活函数 (Activation function) $f(z)$ 映射出该神经元的输出, 激活函数使 M-P 神经元模型具有非线性信号处理能力。

生物神经元中的激活函数是阶跃函数, 如图 2-2(a) 所示。阶跃函数的输出是“0”或“1”, 分别表示神经元处于“抑制”或“兴奋”状态, 用公式表达为:

$$f(x) = \text{sgn}(x) = \begin{cases} 1, & x \geq 0; \\ 0, & x < 0. \end{cases} \quad (2-2)$$

然而, 阶跃函数的输出只有两个状态, 极大地限制了神经元的表达能力, 因此人们设计了平滑且连续的 Sigmoid 函数^[39], 如图 2-2(b) 所示。Sigmoid 函数表达为:

$$f(x) = \text{sigmoid}(x) = \frac{1}{1 + e^{-x}} \quad (2-3)$$

阶跃函数和 Sigmoid 函数有共同点也有区分点。它们的共同点在于, 当输入值较小时, 它们将输出“0”或接近于“0”, 当输入值较大时, 他们将输出“1”或者接近于“1”。它们的区别在于, 阶跃函数的输出是 0/1 的二元信号, 而 Sigmoid 函数输出的是连续的流动信号。并且, Sigmoid 函数处处可导, 导数非常简洁, 这使得 Sigmoid 函数在反向传播的计算中非常方便, 其导数为:

$$\text{sigmoid}'(x) = \text{sigmoid}(x) * (1 - \text{sigmoid}(x)) \quad (2-4)$$

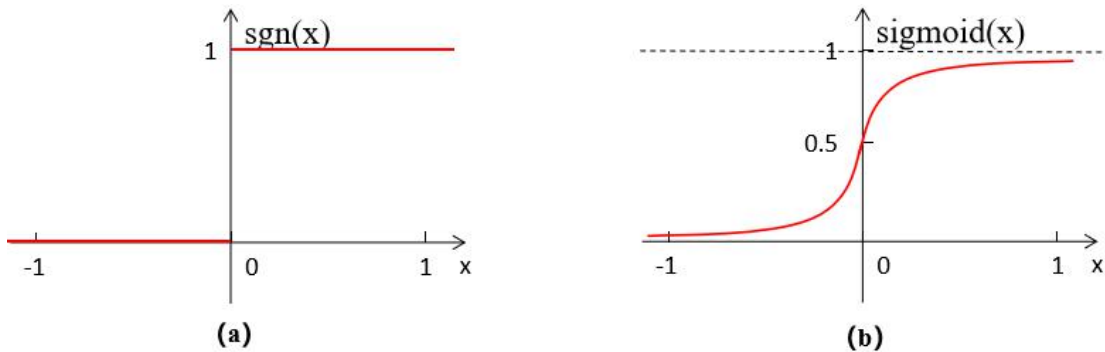


图 2-2 典型的激活函数; (a) 阶跃函数; (b) Sigmoid 函数

Figure 2-2. Typical activation function, (a) Step function, (b) Sigmoid function

2.1.2 多层感知机

由多个神经元嵌套链接而成的模型叫做神经网络，多层感知机是一种经典的神经网络结构。多层感知机由三部分组成：输入层、隐含层、输出层。其中，隐含层是输入层和输出层中间的神经元，特别地，包含一层隐含层的感知机称为单隐层感知机，包含两层隐含层的感知机称为双隐层感知机，图 2-3 是一个单隐层感知机。

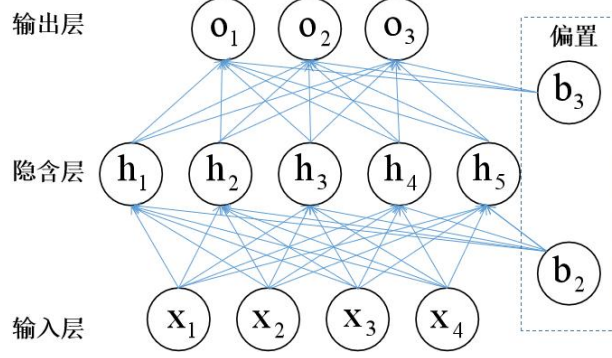


图 2-3 单隐层感知机

Figure 2-3. Single hidden layer perceptron

在这个单隐层感知机中，输入层、隐含层和输出层分别包含 3、5、4 个神经元。其中，输入层只负责传递输入信号，不涉及计算。隐含层和输出层的神经元都属于 M-P 神经元模型，并且都拥有激活函数。 x_i 、 h_i 、 o_i 分别表示输入层、隐含层、输出层中第 i 个神经元的输出信号。隐含层和输出层神经元的计算公式如下：

$$\begin{aligned} h_i &= f(w_2^{i1}x_1 + w_2^{i2}x_2 + w_2^{i3}x_3 + w_2^{i4}x_4 + b_2^i) \\ &= f\left(\sum_{j=1}^4 w_2^{ij}x_j + b_2^i\right) \end{aligned} \quad (2-5)$$

$$\begin{aligned} o_i &= f(w_3^{i1}h_1 + w_3^{i2}h_2 + w_3^{i3}h_3 + w_3^{i4}h_4 + w_3^{i5}h_5 + b_3^i) \\ &= f\left(\sum_{j=1}^5 w_3^{ij}h_j + b_3^i\right) \end{aligned} \quad (2-6)$$

其中， w_l^{ij} 表示第 l 层的第 i 个神经元与上一层中第 j 个神经元的连接权重， b_l^i 表示第 l 层中第 i 个神经元的偏置， f 表示激活函数。

上述公式也可以用矩阵形式描述：

$$\begin{aligned} \mathbf{H} = \begin{bmatrix} h_1 \\ h_2 \\ h_3 \\ h_4 \\ h_5 \end{bmatrix} &= f\left(\begin{bmatrix} w_2^{11} & w_2^{12} & w_2^{13} & w_2^{14} \\ w_2^{21} & w_2^{22} & w_2^{23} & w_2^{24} \\ w_2^{31} & w_2^{32} & w_2^{33} & w_2^{34} \\ w_2^{41} & w_2^{42} & w_2^{43} & w_2^{44} \\ w_2^{51} & w_2^{52} & w_2^{53} & w_2^{54} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} + \begin{bmatrix} b_2^1 \\ b_2^2 \\ b_2^3 \\ b_2^4 \\ b_2^5 \end{bmatrix}\right) \\ &= f(\mathbf{W}_2 \mathbf{X} + \mathbf{B}_2) \end{aligned} \quad (2-7)$$

$$\begin{aligned} \mathbf{O} = \begin{bmatrix} o_1 \\ o_2 \\ o_3 \end{bmatrix} &= f \left(\begin{bmatrix} w_3^{11} & w_3^{12} & w_3^{13} & w_3^{14} & w_3^{15} \\ w_3^{21} & w_3^{22} & w_3^{23} & w_3^{24} & w_3^{25} \\ w_3^{31} & w_3^{32} & w_3^{33} & w_3^{34} & w_3^{35} \end{bmatrix} \begin{bmatrix} h_1 \\ h_2 \\ h_3 \\ h_4 \\ h_5 \end{bmatrix} + \begin{bmatrix} b_3^1 \\ b_3^2 \\ b_3^3 \\ b_3^4 \\ b_3^5 \end{bmatrix} \right) \\ &= f(\mathbf{W}_3 \mathbf{H} + \mathbf{B}_3) \end{aligned} \quad (2-8)$$

2.1.3 反向传播算法

输入信号经过神经网络前向传递产生输出信号，输出信号与期望输出值存在误差，通过反向传播算法（error BackPropagation algorithm, BP 算法）更新权重矩阵，训练神经网络产生期望的输出。

以图 2-3 中的单隐层感知机为例，对反向传播算法进行演算。该网络的输出为 o_i ，真实标签为 o_i^t ，则平方差损失函数 E 定义为：

$$E = \frac{1}{2} \sum_{i=1}^n (o_i^t - o_i)^2 \quad (2-9)$$

其中， n 表示输出层中神经元的个数。反向传播算法通过迭代的方式逐步修正神经元中的权重 w_l^{ij} 和偏置 b_l^i ，从而最小化该模型在训练集上的累计平方差损失值。具体地，反向传播算法使用梯度下降策略，对权重或者偏置在负梯度方向进行调整，定义学习率为 η ，每次调整的值分别为：

$$\Delta w_l^{ij} = -\eta \frac{\partial E}{\partial w_l^{ij}} \quad (2-10)$$

$$\Delta b_l^i = -\eta \frac{\partial E}{\partial b_l^i} \quad (2-11)$$

调整后的值为：

$$v = v + \Delta v \quad (2-12)$$

分别对输出层中神经元的权重 w_3^{ij} 和偏置 b_3^i 进行求导：

$$\frac{\partial E}{\partial w_3^{ij}} = \frac{\partial E}{\partial o_i} \cdot \frac{\partial o_i}{\partial \beta_3^i} \cdot \frac{\partial \beta_3^i}{\partial w_3^{ij}} \quad (2-13)$$

$$\frac{\partial E}{\partial b_3^i} = \frac{\partial E}{\partial o_i} \cdot \frac{\partial o_i}{\partial \beta_3^i} \cdot \frac{\partial \beta_3^i}{\partial b_3^i} \quad (2-14)$$

其中 β_3^i 表示输出层中第 i 个神经元的输入，经过 Sigmoid 激活函数映射后得到输出 o_i ，即 $o_i = f(\beta_3^i)$ 。 β_3^i 可表示为：

$$\beta_3^i = \sum_{j=1}^m w_3^{ij} h_j + b_3^i \quad (2-15)$$

因此式(2-13)和式(2-15)可以进一步表示为:

$$\frac{\partial E}{\partial w_3^{ij}} = \frac{\partial E}{\partial o_i} \cdot o_i' \cdot h_j = (o_i - o_i^t) \cdot o_i \cdot (1 - o_i) \cdot h_j \quad (2-16)$$

$$\frac{\partial E}{\partial b_3^i} = \frac{\partial E}{\partial o_i} \cdot o_i' = (o_i - o_i^t) \cdot o_i \cdot (1 - o_i) \quad (2-17)$$

经过一次迭代, w_3^{ij} 和 b_3^i 的值更新为:

$$w_3^{ij} = w_3^{ij} - \eta \cdot \frac{\partial E}{\partial w_3^{ij}} \quad (2-18)$$

$$b_3^i = b_3^i - \eta \cdot \frac{\partial E}{\partial b_3^i} \quad (2-19)$$

同理, 对于其他层的权重和偏置参数都可以使用链式法则求得导数, 再根据式(2-12)进行更新。

2.2 卷积神经网络

卷积神经网络 (Convolutional Neural Network, CNN) 是一种特殊的神经网络结构, 通过组合卷积层、激活层、池化层和全连接层从原始数据中学习特征^[40]。卷积神经网络是深度学习中最常用的特征学习方法, 被广泛用于图像识别、自然语言处理和语音识别中。

2.2.1 卷积层

卷积层 (Convolution Layer) 通过卷积核与输入数据矩阵进行点乘求和的操作来提取图像特征, 一个卷积神经网络中往往嵌套多层卷积层, 低层卷积层提取纹理、颜色等简单特征, 深层卷积层提取更深的语义特征。卷积层具有权值共享、局部连接的特点, 这两个特点降低了参数数量, 防止网络过拟合。权值共享指对于输入特征图中的每一个位置, 卷积核的权重参数是共享的, 相比于全连接层网络, 权值共享的特点减少了卷积层中参数的个数, 使运算变得简洁而高效。所谓局部连接, 就是卷积核在与输入特征图进行点乘求和的计算时, 每次卷积核只与特征图的局部节点连接。局部连接的好处在于减少参数个数, 此外, 图像本身的局部相关性比较强, 图像中距离较远的像素点关联不大, 因此局部连接也有助于提取图像的局部特征。

卷积层的计算过程可分为两种: 无填充的卷积过程和带填充的卷积过程。图 2-4 演示了无填充的卷积过程, $X_1 \sim X_n$ 表示 $5 \times 5 \times N$ 的输入特征图, $W_1 \sim W_n$ 表示一个 $3 \times 3 \times N$ 的卷积核, 也就是该卷积神经元的权重参数矩阵, Bias 表示该卷积核的偏置, Y 表示输出特征图。卷积核在特征图上进行步长为 1 的卷积计算公式如下:

$$y_{ij} = \sum_{m=1}^n \sum_{h=i}^{i+2} \sum_{q=j}^{j+2} x_{hq}^m \cdot w_{hq}^m + b \quad (2-20)$$

卷积核在特征图上进行滑动，每次滑动过程中，卷积核连接特征图的局部节点，每个通道的卷积核矩阵与对应特征图进行点乘求和，并将所有通道的求和值累加，再加上偏置即可得到该卷积层的输出特征图。需要注意的是，这里只用了一个卷积核，因此输出的特征图是单通道的，通常一个卷积层会使用多个卷积核，每个卷积核输出一层特征图，最终的特征图是一个三维的数组。

无填充的卷积层会使输出特征图尺寸变小，如图 2-4 所示，输入特征图的尺寸是 $5 \times 5 \times N$ ，而输出特征图的尺寸是 $3 \times 3 \times N$ 。特征图尺寸缩小的原因是因为输入特征图的边缘不能与卷积核的中心连接，因此，研究人员将输入特征图的四周边缘分别扩充 $(k-1)/2$ 层节点， k 表示卷积核的宽度，填充节点的值往往使用 0 或者相邻节点的数值。

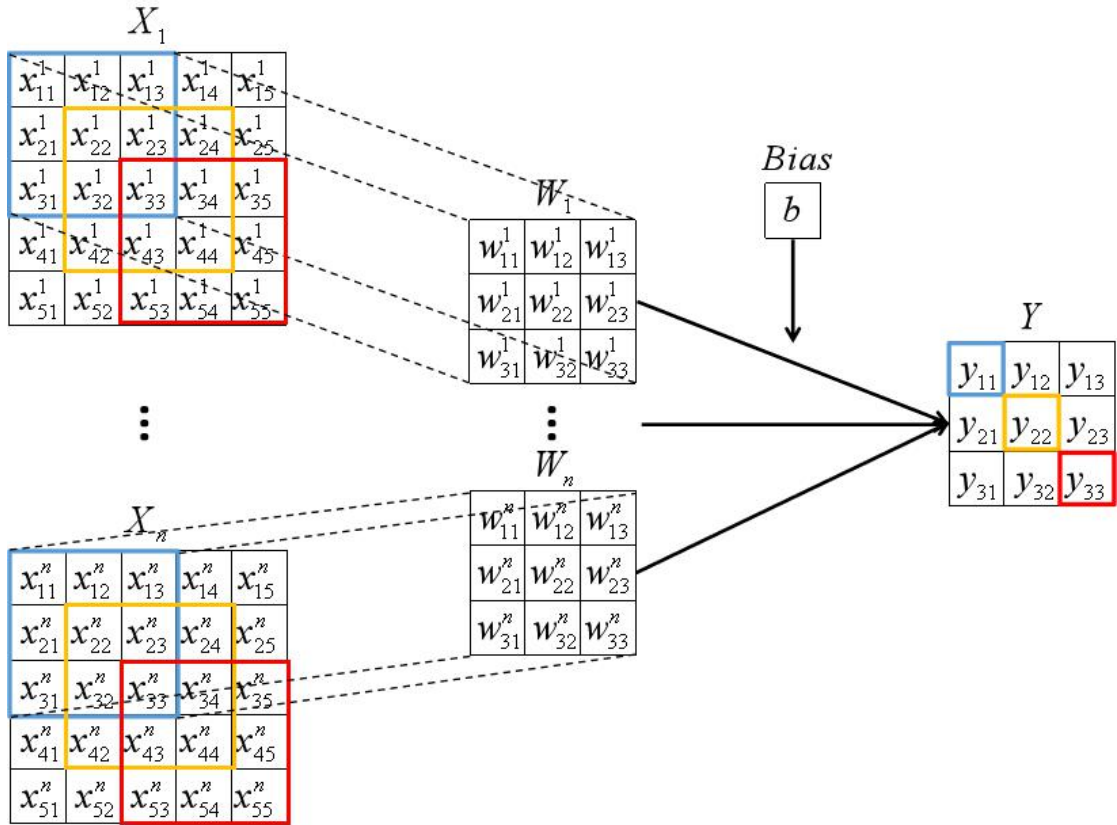


图 2-4 无填充的卷积过程

Figure 2-4. Convolution process without padding

2.2.2 激活层

激活层（Activation Layer）的作用是对卷积层的输出特征图进行非线性映射，使卷积神经网络具备非线性能力。激活层通常由激活函数组成，对于卷积

层的输出特征图 Y ，激活层的输出为 $A = f(Y)$ ， f 表示激活函数。典型的激活函数有阶跃函数和 Sigmoid 函数，在 2.1.1 节中已有介绍。然而，阶跃函数的输出只有 0/1，而 Sigmoid 饱和区导数太小，使得在反向传播时参数变化太慢，导致 Sigmoid 函数在深层网络中存在梯度消失问题。因此，现有的卷积神经网络中往往使用 ReLU 函数：

$$f(x) = \begin{cases} x, & x \geq 0; \\ 0, & x < 0. \end{cases} \quad (2-21)$$

ReLU 函数的好处是使得求导计算更方便，同时在一定程度上缓解了梯度消失的问题，然而，ReLU 使一部分输出为 0，虽然使网络具有一定的稀疏性，但是有可能出现神经元“坏死”的情况，即该神经元的输出一直为 0。He 等人^[41]提出了 ReLU 函数的变体 PReLU 函数，其公式如下：

$$f(x) = \begin{cases} x, & x \geq 0; \\ a_i x, & x < 0. \end{cases} \quad (2-22)$$

其中， a_i 表示 PReLU 函数在特征图第 i 个通道的负向斜率。相比于 ReLU，PReLU 避免神经元“坏死”的情况发生，减小了过拟合的风险。ReLU 和 PReLU 的图形如图 2-5 所示。

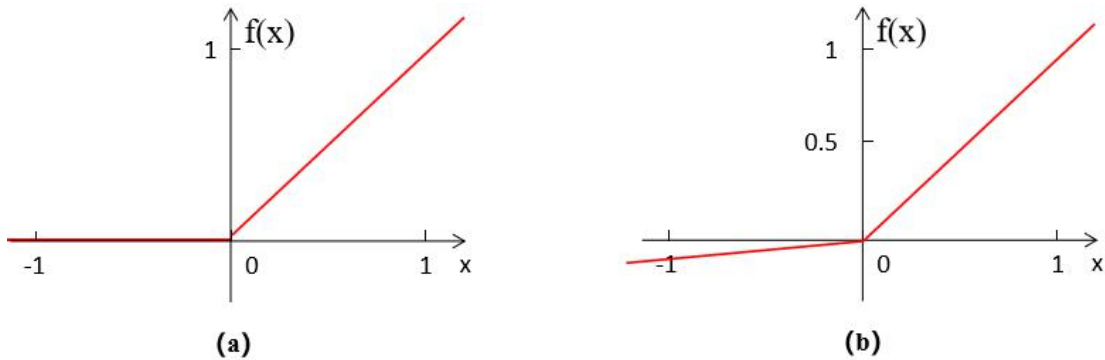


图 2-5 ReLU 和 PReLU 激活函数；(a) ReLU 函数；(b) PReLU 函数

Figure 2-5. Activation function of ReLU and PReLU, (a) ReLU function, (b) PReLU function

2.2.3 池化层

池化层（Pooling Layer）是对特征图中的信息进行筛选，提取更有价值的特征。常用的池化层有：最大值池化、均值池化。图 2-6(a)表示最大值池化的过程，它在一个 4×4 的特征图上使用 2×2 的滤波器以步长为 2 进行滑动扫描，将扫描范围内的最大值输出。最大值池化是最常用的池化操作，有助于学习到图像的纹理和边缘特征。图 2-6(b)表示均值池化的过程，扫描的过程与最大值池化一致，不同的是，均值池化的输出值是扫描范围内特征图节点的均值。

池化层有三方面好处：第一，筛选掉特征图中的非代表性信息，从而降低信息冗余；第二，特征图减小之后，可以减少网络后面连接的参数，尤其是全连接层神经元包含的参数，防止过拟合；第三，当输入图片发生平移或者旋转时，池化提取局部位置中的显著信息，使特征图尽可能保持不变，提升了模型的尺度不变性和旋转不变性。

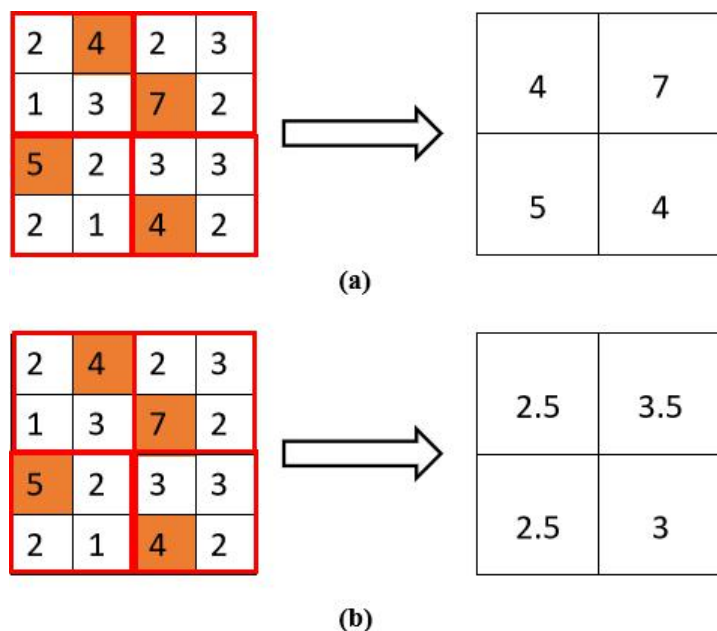


图 2-6 两种典型的池化方式：(a) 最大值池化；(b) 均值池化

Figure 2-6. Two typical pooling methods, (a) Max Pooling, (b) Average Pooling

2.2.4 全连接层

全连接层（Full Connected Layer）中的神经元连接上一层中所有的神经元，如图 2-7 所示。全连接层往往出现在卷积神经网络中最后几层，用于融合全局信息，将上层信息进行汇总。需要注意的是，全连接层的参数非常多，参数数量在整个卷积神经网络中占比非常大，因此不宜添加太多全连接层。

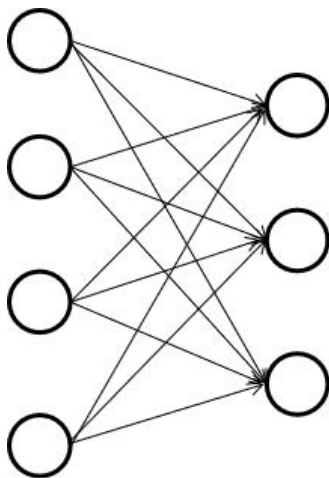


图 2-7 全连接层

Figure 2-7. Full Connected Layer

2.3 行人重识别方法

2.3.1 基于全局特征的 Re-ID 方法

基于全局特征的行人重识别方法利用卷积神经网络强大的特征提取能力，将输入行人图像作为一个整体进行特征提取，即利用整图得到一个特征向量。基于全局特征的行人重识别的方法往往集中在改进网络结构和设计损失函数。早期全局特征网络直接采用 ResNet50 等在分类任务中表现优异的神经网络，后期研究人员主要针对行人重识别的特点进行改进，如 Res-Mid^[42]以 ResNet50 为骨干网络，融合中层特征和高层特征来提升特征表达能力。基于全局特征损失函数主要提升网络对行人全局特征的提取，而提取细节特征的能力较弱。

2.3.2 基于局部特征的 Re-ID 方法

基于局部特征的行人重识别方法将行人图像分割成局部片段，再分别进行特征提取和匹配，相比于全局特征的方法，基于局部特征的行人重识别方法可以关注到行人的细节特征。其分割方式往往分为两种：水平分割和骨骼关键点分割。基于水平分割的行人重识别算法有 PCB、MGN 等，将行人图像自上而下分割成水平图像块，分别提取每个水平图像块的特征，融合子特征作为最终的行人特征向量。

然而，由于行人姿态、姿势的变化，按照均等分割的水平分割方法可能失效。于是研究人员根据先验知识将行人不同的部位进行对齐，这些先验知识主要基于骨骼关键点，通过预训练的骨骼关键点模型来获得人体不同部位的图像。再对这些子图像分别特征提取的工作，最后融合不同部位的特征。

2.3.3 基于视频序列的 Re-ID 方法

单独帧的行人图像包含信息有限，而基于视频序列的行人重识别方法不仅利用了单帧图像的内容，也利用了上下帧之间的运动信息。联系上下帧的视频序列方法主要结合了卷积神经网络和递归循环网络，首先通过卷积神经网络提取每一帧图像的空间信息，再利用递归循环网络来提取时序特征。

此外，基于视频序列的行人重识别方法也常常利用光流特征来获取行人的运动信息。首先提取相邻两帧的光流特征，再将光流特征和卷积神经网络提取的空间特征进行融合，输入到递归循环网络中提取时序特征。最后再用身份损失、对比损失等损失函数进行训练。

2.4 本章小结

本章介绍了人工神经网络和卷积神经网络的主要组成部分，以及不同类型的行人重识别方法。首先从仿生学的角度介绍了人工神经网络中的基本单元：M-P 神经元模型，进而介绍了由神经元组成的多层感知机结构，同时推导了基

于梯度下降的反向传播算法在神经网络中的计算公式。卷积神经网络是人工神经网络的分支，也是当下深度学习在图像领域的重要工具。本章介绍了卷积神经网络常用的四个结构：卷积层、激活层、池化层和全连接层，分别演示了它们计算过程和计算方法，以及各个层的功能与利弊。本章末尾介绍了不同类型的基于卷积神经网络的行人重识别方法，后文将介绍基于卷积神经网络设计的跨模态行人重识别模型，通过阅读本章，有助于理解后续的内容。

第三章 基于困难五元组的特征提取框架的设计

现有的单模态行人重识别研究已经非常丰富，主要包含针对单模态行人重识别设计的神经网络和损失函数两方面内容。单模态行人重识别和跨模态行人重识别的研究目标都是如何更好地在图像/视频中提取行人特征，区别在于跨模态行人重识别增加考虑了模态变化的因素。因此，单模态行人重识别网络在迁移至跨模态行人重识别任务时存在不能很好发挥原有性能的问题。单模态行人重识别网络经过多年的积累，针对行人的特性设计了大量特征提取方法，因此将单模态行人重识别网络迁移到跨模态任务中是非常有必要的。

针对以上问题，本章设计了一种基于困难五元组的特征提取框架（Hard Pentaplet and Identity Loss Network, HPILN）。首先，设计了一种单模态行人重识别网络的迁移策略。然后，我们提出一种在训练过程中采样困难五元组的方法。同时，我们设计了困难五元组损失函数，该损失函数考虑了模态内的变化和模态间的变化，使单模态行人重识别网络在跨模态行人重识别任务中可以更好地发挥性能。最后，我们将身份损失函数融合到特征提取框架中，进一步提升了精度。

3.1 特征提取框架的设计

本章设计的特征学习框架如图 3-1 所示，该框架包含四个部分：图像输入模块、特征提取模块、特征采样模块、损失函数模块。具体地，在图像输入模块中我们设计了一种训练批次的选择机制，每次从训练集中随机选择 $2PK$ 张图片，即随机选择 P 个行人的 K 张可见光图片和 K 张红外图片，我们将这种训练批次命名为 **cm-batch**（cross-modality batch）。然后，我们将单模态行人重识别网络改造成特征提取模块，利用卷积神经网络自动提取特征的特性来实现跨模态行人图片的特征提取功能，提取到的特征以特征向量的形式存在。接着，本文设计了一种新的损失函数：困难五元组损失函数，这种损失函数需要对生成的特征向量进行组队，因此特征采样模块的作用就是采样出困难五元组。最后，损失函数模块由困难五元组损失和身份损失级联而成，用于监督网络的训练。

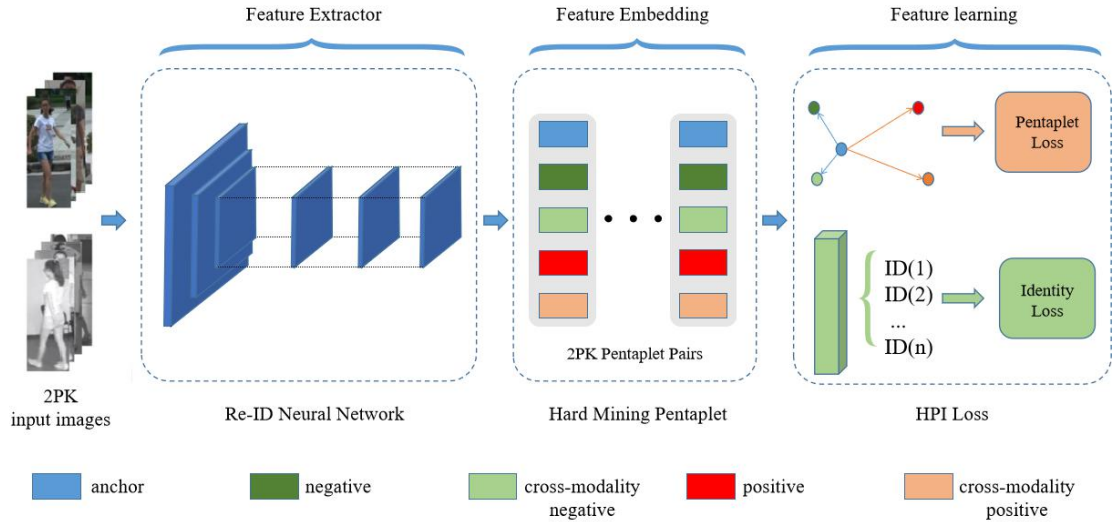


图 3-1 HPILN 特征提取框架

Figure 3-1. The HPILN feature extraction framework

3.1.1 改造单模态行人重识别网络

本文改造现有的单模态行人重识别卷积神经网络，使其在跨模态行人重识别（SYSU-MM01、RegDB）数据集上适用。我们使用在单模态行人重识别数据集上表现优异的卷积神经网络模型作为本框架中的特征提取模块，如 BFE^[43]、Res-Mid^[42]、PCB^[16]、MGN^[17]、MLFN^[44]。

典型的单模态行人重识别卷积神经网络如图 3-2 所示。行人图像经过 CNN 模块得到特征图，特征图转成一维向量后添加一层全连接层（FC-1），FC-1 的输出表示该图像的特征向量。为了使该特征向量具有表征性，通常使用排序损失（Ranking loss）和身份损失（Identity loss）进行约束。具体地，排序损失函数采用度量学习的思想，来减少类内距离和类间距离，常见的排序损失函数有三元组损失、对比损失等。身份损失通常采用交叉熵损失函数，将每个身份的行人归为一类，来减小类内距离。FC-2 是全连接层，其神经元个数为训练集中行人身份的个数，用于身份损失函数的计算。

我们将改造的部分在图 3-2 中用红色表示，主要包括 Ranking loss 和 FC-2 全连接层两个部分。本章设计了一种新的 Ranking loss：困难五元组损失（hard pentaplet loss），在 3.1.2 节中详细介绍了该损失函数。FC-2 主要修改了神经元的个数，该神经元个数等于训练集中的行人身份数量。因此，我们在对 SYSU-MM01 和 RegDB 数据集进行训练时，分别将 FC-2 的神经元个数设置为 395 和 206 个。

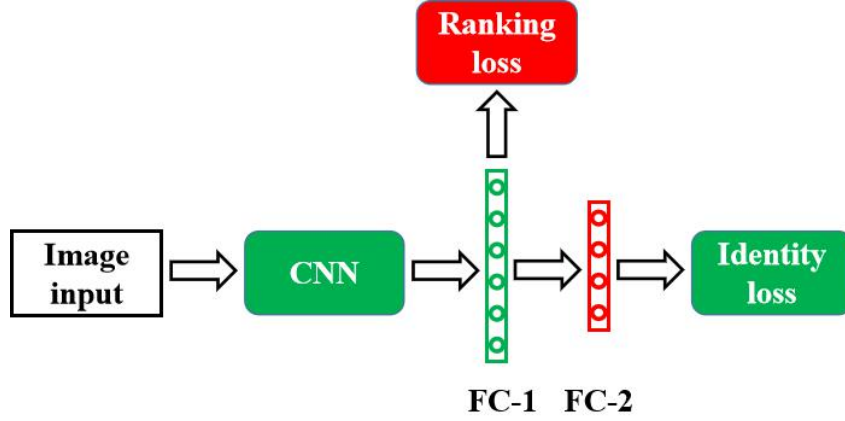


图 3-2 典型的单模态行人重识别网络

Figure 3-2. A typical single modality Re-ID CNN model

本章设计的特征提取框架与典型的单模态行人重识别网络在组成模块上非常相似，目的是方便将来性能优异的单模态行人重识别网络能够很方便的使用 HPILN 框架迁移到跨模态行人重识别任务中，这对未来跨模态行人重识别领域的发展有一定的帮助。图 3-1 中“Feature Extractor”模块对应图 3-2 中的 CNN 模块，我们采用单模态行人重识别中设计的 CNN 作为跨模态行人重识别任务中的特征提取模块。由于单模态行人重识别网络在对行人的特征提取做了大量的设计，相比于分类任务中的卷积神经网络，专用的行人重识别网络能够从行人图像中提取更有表征性的特征向量。此外，图 3-1 中“Feature Embedding”模块中的特征向量对应图 3-2 中的 FC-1，图 3-1 中“Feature Learning”模块中的两个损失函数分别对应图 3-2 中的 Ranking loss 和 Identity loss。

3.1.2 困难五元组损失函数

本章设计的困难五元组损失函数改进自困难三元组损失函数，而困难三元组损失函数是三元组损失函数的升级版。因此，为了理解困难五元组损失函数，我们将先简单介绍三元组损失和困难三元组损失函数。

三元组损失函数广泛应用于人脸识别、行人重识别、车辆重识别等图像检索领域。在行人重识别任务中，对于候选三元组 $\{x_i^a, x_i^p, x_i^n\}$ 的锚点图像 x^a (其中 $i \in [1, N]$, N 表示训练批次中三元组的个数)， x^p 表示与 x^a 身份相同的正样本， x^n 表示与 x^a 身份不同的负样本行人图像。使用卷积神经网络作为特征提取器，图像 x 会在欧式空间中被映射为 d 维的特征向量，特征向量可以表示为 $f(x) \in \mathbb{R}^d$ 。度量两张行人图像的相似度可以用欧式距离来度量，其公式如下：

$$d(x_i, x_j) = \|f(x_i) - f(x_j)\|_2 \quad (3-1)$$

三元组损失函数可以表示为：

$$L_{trp} = \sum_i^N [d(x_i^a, x_i^p)^2 - d(x_i^a, x_i^n)^2 + \alpha]_+ \quad (3-2)$$

其中, $[z]_+ = \max(z, 0)$ 。在三元组 $\{x_i^a, x_i^p, x_i^n\}$ 中, $d(x_i^a, x_i^p)$ 表示正样本对 (x_i^a, x_i^p) 的欧式距离, 而 $d(x_i^a, x_i^n)$ 表示负样本对 (x_i^a, x_i^n) 的欧式距离。需要注意的是, 三元组损失函数对欧式距离进行平方处理, 这点将在后续的改进中取消。 α 是一个超参数, 用于迫使正负样本在欧式空间内分开。三元组在三元组损失函数训练结束前后的变化如图 3-3 所示, 正样本对之间的欧式距离将变得更小, 而负样本对之间的欧式距离将变得更大。行人重识别中相似度用欧式距离来表示, 欧式距离越小表示图像越相似, 这意味着三元组损失函数可以使卷积神经网络提取到的特征更有判别性, 即相同行人的特征向量更相似, 不同行人的特征向量更不相似。



图 3-3 三元组损失函数
Figure 3-3. Triplet Loss

Hermans 等人^[22]通过改进三元组的采样方式提出了困难三元组损失函数, 这大大提高了模型训练的速度和精度。在困难三元组损失函数中, 每个训练批次随机选择 P 个身份的行人, 每个行人随机采样 K 张图像, 因此每个批次中包含 PK 张图片。对于该批次中的每一张图片, 在该批次中选取最困难的正负样本对来组成困难三元组。困难正样本表示在该批次中与锚点图像的欧式距离最大的相同身份的图像, 而困难负样本表示在该批次中与锚点图像的欧式距离最小的不同身份的图像。困难三元组可以表示为:

$$L_{trp} = \sum_{i=1}^P \sum_{a=1}^K [\max(d(x_i^a, x_i^p)) - \min(d(x_i^a, x_i^n)) + \alpha]_+ \quad (3-3)$$

其中, $p, n \in [1, K], j \in [1, P], j \neq i$ 。其余参数与式(3-2)中表示的相同。

困难三元组损失函数有效地减小了类内的距离、增大了类间的距离, 而跨模态行人重识别中模态的变化往往更大。困难五元组损失函数在困难三元组损失的基础上额外考虑了模态的变化, 本章设计的困难五元组损失函数不仅兼顾了困难三元组损失函数在处理模态内变化时的优点, 还可以减小模态间变化引起的识别干扰。困难五元组损失函数由困难全局三元组 (hard global triplet, HGT) 和困难跨模态三元组 (hard cross-modality triplet, HCT) 损失函数组成。

困难全局三元组损失函数的作用是同时处理模态内的变化和模态间的变化。随机选取 P 个身份的行人，每个行人随机选取 K 张可见光图像和 K 张红外图像，组成一个训练批次（cm-batch）。遍历训练批次中的每一张图像作为锚点 x^a ，跨模态负样本集和 x^{cn} 同模态负样本集和 x^{in} 的总和组成全局负样本集和 x^{gn} ，而跨模态正样本集和 x^{cp} 同模态正样本集和 x^{ip} 的总和组成全局正样本集和 x^{gp} 。困难全局三元组的公式如下：

$$L_{hgt} = \sum_{i=1}^P \sum_{a=1}^{2K} [\max(d(x_i^a, x_i^{gp})) - \min(d(x_i^a, x_i^{gn})) + \alpha]_+ \quad (3-4)$$

其中， $p, n \in [1, 2K], j \in [1, P], j \neq i, p \neq a$ 。其余参数与式(3-2)中表示的相同。

尽管困难全局三元组损失可以同时处理模态内变化和模态间变化，但是在跨模态行人重识别中模态间变化往往更大。因此我们设计了困难跨模态三元组损失函数，其公式表达为：

$$L_{hct} = \sum_{i=1}^P \sum_{a=1}^{2K} [\max(d(x_i^a, x_i^{cp})) - \min(d(x_i^a, x_i^{cn})) + \alpha]_+ \quad (3-5)$$

其中，当 $a \geq K$ 时， $cp, cn \in [1, K]$ ，否则， $cp, cn \in [K+1, 2K]$ 。此外， $j, k \in [1, K], k \neq i$ 。

我们设计的困难五元组损失函数由困难全局三元组和困难跨模态三元组损失函数组成。在每个训练批次中，我们将采样的困难全局三元组集和 $\{x_i^a, x_j^{gp}, x_k^{gn}\}$ 和困难跨模态三元组集和 $\{x_i^a, x_h^{cp}, x_t^{cn}\}$ 组成困难五元组集和 $\{x_i^a, x_j^{gp}, x_k^{gn}, x_h^{cp}, x_t^{cn}\}$ 。其中 x_j^{gp} 和 x_h^{cp} 、 x_k^{gn} 和 x_t^{cn} 有可能是同一张图片。困难五元组损失函数的公式如下：

$$L_{hp} = \frac{1}{2PK} (L_{hgt} + L_{hct}) \quad (3-6)$$

困难五元组损失函数的作用是在相同模态和跨模态情景下减小类内距离、增大类间距离，如图 3-4 所示。此外，困难五元组的采样方式使用有限的训练集图像生成了大量的困难五元组，丰富了训练样本，同时加速了模型的训练。

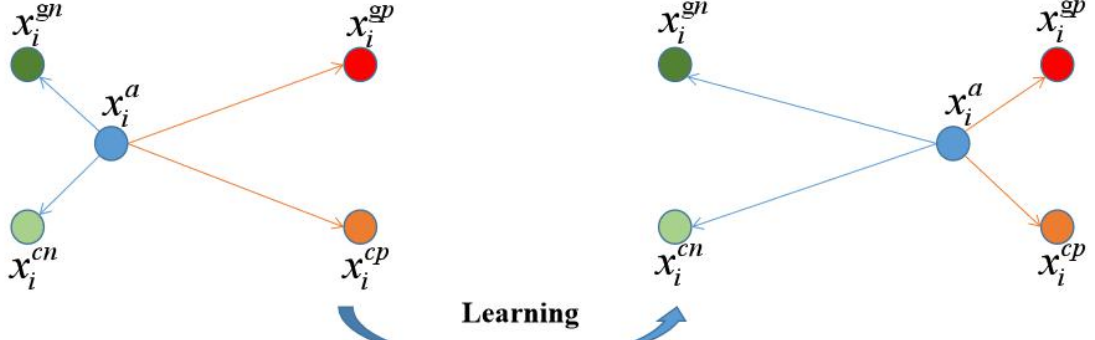


图 3-4 困难五元组损失函数

Figure 3-4. Hard pentaplet loss

3.1.3 级联身份损失

身份损失函数可以提取不同类别的身份信息，因此我们使用身份损失去提升模型处理类内变化的能力。忽视模态的变化，我们把不同模态中的同一行人归为同一类别，身份损失函数用 softmax 损失来表示：

$$L_{id} = \frac{1}{2PK} \sum_{i=1}^{2PK} -\log \left(\frac{e^{f_{y_i}}}{\sum_j e^{f_{y_j}}} \right) \quad (3-7)$$

其中， $2PK$ 表示训练批次中的图片总数， f 是最后一层全连接层 FC-2 的输出向量， $f_j, j \in [1, H]$ 表示第 j 个神经元的输出值，而 H 表示训练集中行人类别的个数。 y_i 表示输入图片 x_i 的真实标签，因此 f_{y_i} 表示输入图片 x_i 的类别得分。

在本章设计的 HPILN 框架的损失函数模块中，我们级联了困难五元组损失函数与身份损失函数，级联后的损失函数我们称之为 HPI（Hard Pentaplet and Identity）损失函数。HPI 损失函数的公式如下：

$$L_{HPI} = L_{hp} + L_{id} \quad (3-8)$$

3.2 实验与分析

3.2.1 实验设置与环境

本章在 SYSU-MM01 跨模态行人重识别数据集上测试了 HPILN 框架的性能。使用 CMC 曲线和 mAP 平均检索精度这两种评测指标来检测性能，其评测过程遵守数据集官方的标准，在 1.3.2 节中已详细描述。在实验环境方面，我们使用 NVIDIA GeForce 1080Ti 显卡和 Pytorch 深度学习框架作为硬件和软件平台。

我们选择近三年在单模态行人重识别任务中表现良好的网络结构作为 HPILN 框架中的特征提取模块，具体选择了 BFE^[43]、Res-Mid^[42]、PCB^[16]、

MGN^[17]、MLFN^[44]等网络。由于网络结构不同，因此其图片输入尺寸和特征向量的维度也不同，针对不同的网络，我们遵守原本的设计方案设置图片输入尺寸和特征向量维度，具体参数如表 3-1 所示。在反向传播算法中，我们设置初始学习率为 $3 * 10^{-4}$ ，并采用自适应矩估计(Adam)优化器^[45]来动态调整学习率。每个批次中我们随机选取 P 个行人的 K 张可见光图片和 K 张红外图片，合计为 $2PK$ 张图片。实验中我们将 P 设置为 8， K 设置为 4，因此训练批次大小设置为 64。此外，我们对每张输入图片进行随机裁剪和水平翻转的策略进行数据增强。困难五元组损失函数中的超参数 α 在[0.3,0.6,0.9,1.2,1.5,1.8]范围内逐个取值进行实验。

表 3-1 图片输入尺寸和特征维度
Table 3-1. Picture input size and feature dimensions

模型	宽度*高度	特征维度
<i>Res-Mid</i>	224*224	3072
<i>MGN</i>	128*384	2048
<i>PCB</i>	224*224	12288
<i>BFE</i>	128*256	1024
<i>MLFN</i>	224*224	1024

3.2.2 实验对比

为了验证 HPILN 框架性能的优越性，我们对比了近三年在顶会或顶刊中发表的方法：zero-padding^[6]、cmGAN^[34]、eBDTR^[26]、D-HSME^[46]、IPVT^[47]、D2RL^[37]。SYSU-MM01 数据集包含四种实验模式：All-search Single-shot (ASSS)、All-search Multi-shot (ASMS)、Indoor-search Single-shot (ISSS)、Indoor-search Multi-shot (ISMS)，我们对这四种都进行了测试。表 3-2 展示了本章设计的 HPILN 框架在 ASSS 模式下的实验结果，结果显示我们的五种方法 (Res-Mid+HPI、MGN+HPI、PCB+HPI、MLFN+HPI、BFE+HPI) 的 Rank1、Rank10、mAP 精度都高于其他方法。尤其是使用困难五元组级联身份损失函数 (HPI Loss) 训练的 BFE 模型，在 ASSS 验证模式下的 Rank1 高于现有最高精度的 12.46%。

为了更全面地测试 HPILN 的性能，本文在 SYSU-MM01 数据集的另外三种验证模式 (ASMS、ISSS、ISMS) 中测试 HPILN 框架，实验结果如表 3-3 所示，加粗部分表示精度最高。

可以看出，HPILN 框架在 SYSU-MM01 数据集的四种验证模式中都表现良好，体现了该框架的优越性、鲁棒性。

表 3-2 HPILN 在 ASSS 模式下的实验结果

Table 3-2. HPILN experimental results in ASSS mode

方法	Rank1	Rank10	mAP
Zero-padding	14.80	54.12	15.95
cmGAN	26.97	67.51	27.80
eBDTR	27.82	67.34	28.42
D-HSME	20.68	62.74	23.12
IPVT	23.18	51.21	22.49
D2RL	28.9	70.6	29.2
Res-Mid+HPI	40.49	83.61	41.64
MGN+HPI	39.77	79.78	41.12
PCB+HPI	33.29	80.66	35.15
MLFN+HPI	33.34	78.54	36.13
BFE+HPI	41.36	84.78	42.95

表 3-3 HPILN 框架在 ASMS、ISSS、ISMS 模式下的实验结果

Table 3-3. HPILN experimental results in ASMS、ISSS、ISMS mode

方法	ASMS Rank1/mAP	ISSS Rank1/mAP	ISMS Rank1/mAP
Res-Mid+HPI	47.70/35.15	45.65/56.19	50.79/46.21
MGN+HPI	44.86/34.88	44.06/54.52	50.55/44.90
PCB+HPI	38.55/28.16	39.70/50.49	46.86/40.93
MLFN+HPI	39.45/29.52	36.25/47.99	41.99/38.43
BFE+HPI	47.56/ 36.08	45.77/56.52	53.05/47.48

3.2.3 分析级联损失函数的作用

为了验证级联困难五元组损失和身份损失的作用，我们在网络结构不变的情况下，修改损失函数的种类，以此来验证级联损失函数的优越性能。在该实验中，我们选取了身份损失（Identity Loss）、中心损失（Center Loss）、困难三元组损失（HT Loss）、困难三元组损失级联身份损失（HTI Loss）、困难五元组损失（HP Loss）、困难五元组损失级联身份损失（HPI Loss）来对比，我们在五种模型上分别添加上述损失函数，我们将 Rank1 和 mAP 的实验结果绘制成柱状图，如图 3-5 所示。显然，困难五元组损失（HP Loss）效果优于身份损失、中心损失、困难三元组损失，在 PCB 网络中表现得尤为明显，我们设计的 HP 损失和 HPI 损失几乎高出其他损失函数一倍，这表示我们设计的困难五元组

损失在跨模态行人重识别任务中比一些经典的损失函数更适用。并且，这也说明了困难五元组损失级联身份损失（HPI Loss）能进一步提升 HP Loss 的性能。

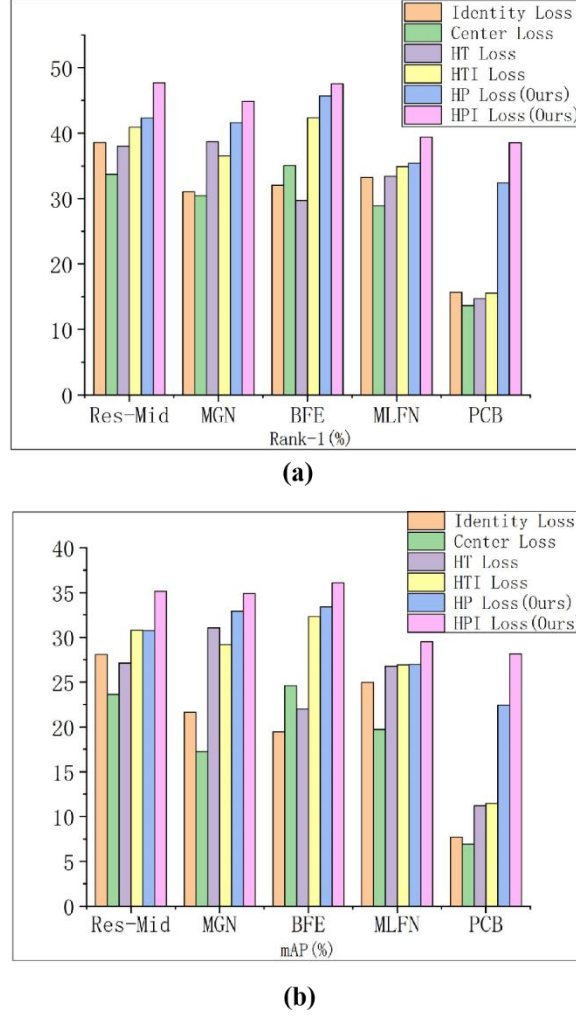


图 3-5 损失函数的对比结果；(a) Rank-1；(b) mAP
Figure 3-5. Comparison of loss functions, (a) Rank-1, (b) mAP

此外，为了更全面地展示不同 Rank 指标中的精度，我们绘制了不同损失函数训练的模型的 CMC 曲线，如图 3-6 所示。该图展示了 Rank 精度随位列的变化，蓝色曲线和红色曲线分别代表我们设计的 HP Loss 和 HPI Loss。显然，在不同网络中，我们设计的 HP Loss 和 HPI Loss 在不同的 Rank 指标中都取得了领先的精度。

为了探究不同损失函数下的特征分布，我们做了特征可视化的实验。具体地，我们把 Res-Mid 的 FC-1 全连接层神经元个数修改为 2，从而学习到二维特征，并把二维特征绘制到平面坐标系中，其结果如图 3-7 所示，五种颜色表示五个不同身份的行人，圆型和星型点分别表示可见光图像和红外图像的特征分布点。如图 3-7 (a)所示，当只用身份损失训练模型时，生成的特征能粗略的把不同的类区分开来，但是特征分布不是很有判别性，原因在于身份损失只关注

减小类内距离，导致类间距离没有很好的被缩小。图 3-7 (b)展示了 HP Loss 训练下的特征分布，类与类之间存在较大的裕度，并且每一类都聚集在簇内，这表明 HP Loss 既可以减小类内距离，又可以增大类间距。图 3-7 (c)是级联损失 HPI Loss 的特征可视化结果，它聚集了前两个损失函数的优点，类内距离更小、类间距离更大。并且相同行人在不同模态中的特征也更聚合，这表示 HPI Loss 可以同时处理模态内变化和模态间变化，并且效果显著。

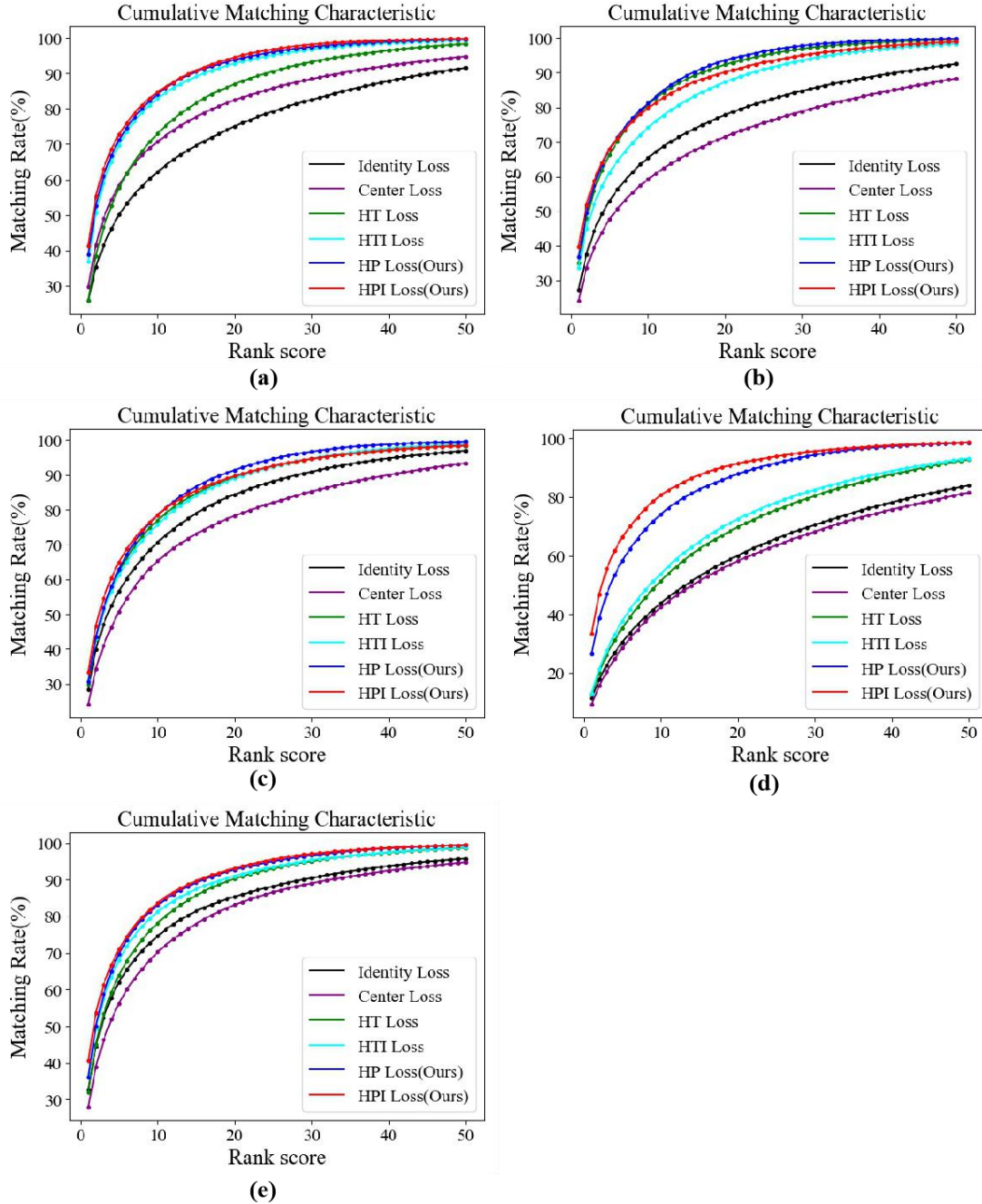


图 3-6 不同损失函数的 CMC 曲线; (a) BFE; (b) MGN; (c) MLFN; (d) PCB; (e) Res-Mid

Figure 3-6. CMC curves of different loss functions, (a) BFE, (b) MGN, (c) MLFN, (d) PCB, (e) Res-Mid

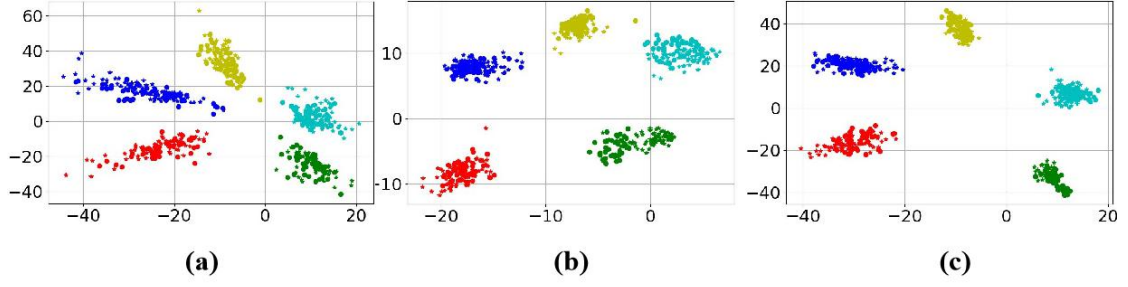


图 3-7 损失函数特征可视化；(a) 身份损失；(b) 困难五元组损失；(c) 困难五元组级联身份损失

Figure 3-7. Visualization of loss function characteristics, (a) Identity Loss, (b) HP Loss, (c) HPI Loss

3.2.4 卷积特征可视化

为了说明我们模型的有效性，我们使用 Grad-CAM 方法[48]来获取卷积特征激活图。具体地，我们一共进行 15 个实验，分别使用 HPI、HTI、ID loss 对 PCB、BFE、MLFN、Res-Mid、MGN 网络进行训练，得到 15 个模型。然后在卷积神经网络前向传播时获取最后一层卷积层的特征图，并对特征图在通道维度进行加权求和，得到的激活图与原图进行叠加，得到的激活图如图 3-8 所示，颜色越鲜艳表示响应越强烈。

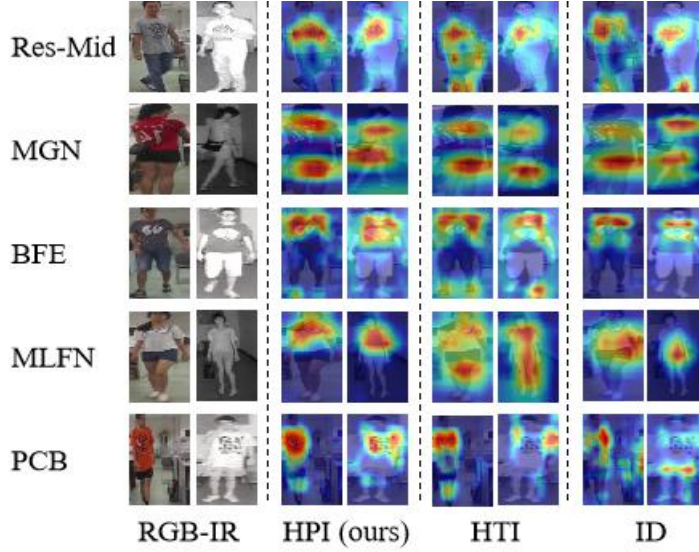


图 3-8 卷积特征激活图

Figure 3-8. Convolution feature activation map

我们发现模型对行人的身体有强烈的响应，而不关注行人背景区域。特别地，模型对行人身体上的显著性区域有更强烈的响应，比如衣服上的图标。我们对比了不同损失函数（HPI、HTI、ID Loss）的卷积特征可视化结果，发现本章提出的 HPI Loss 可以更好地关注到不同模态的相同行人的显著性区域，这意味着 HPI Loss 可以减少模态的影响，从而提取到更具有判别性的特征。

此外，从特征可视化的结果来看，可见光图像和红外图像的共享特征更容易被模型关注，而颜色等模态专属特征容易被模型忽略。

3.3 本章小结

针对跨模态行人重识别任务，本章设计了一种新的特征学习框架(HPILN)。该框架的创新点有：（1）首次将单模态行人重识别网络迁移到跨模态行人重识别任务中，并证明了该策略的优越性；（2）设计了一种新的困难五元组损失函数，该损失函数可以同时处理模态间变化和模态内变化，并且证明了级联身份损失达到了更好的效果。

HPILN 框架的优势在跨模态行人重识别中非常明显。首先，HPILN 提出迁移单模态行人重识别网络到跨模态行人重识别任务中，由于单模态行人重识别已经对提取行人特征进行了大量设计，该策略大大降低了研究人员在设计跨模态行人重识别的特征提取模块时遇到的困难，为未来跨模态行人重识别特征提取器的设计提出新的解决方案。其次，本章设计的困难五元组损失函数既可以同时处理模态间变化和模态内变化，又可以减小类内距离、增大类间距离，困难采样策略有效地提升了网络训练的收敛速度。更重要的是，困难五元组损失级联身份损失训练的模型远超现有网络的精度，推动了深度学习在跨模态行人重识别中进行可靠应用的可能性。

第四章 基于多粒度特征的双通道网络的设计

上一章介绍了 HPILN 特征提取框架，该框架以单模态行人重识别网络作为特征提取模块，如果将该特征提取模块视为一个整体，那么上一章介绍的 HPILN 框架实际上是一个基于单通道的网络结构。单通道网络结构共用一个 CNN 模块去学习可见光图像和红外图像的共同特征，在一定程度上限制了网络提取异质模态细节特征的能力。

本章设计了一个基于多粒度特征的双通道网络。该网络的特征提取模块分为两个通道，这两个通道在结构上一致，但参数不共享，分别提取可见光图像和红外图像的行人特征。此外，特征提取器采用多粒度特征的形式，能够有效地提取行人不同部位的特征，使模型更好地关注到行人的细节。

4.1 双通道网络结构设计

基于多粒度特征的双通道网络结构如图 4-1 所示。首先，双通道网络的前半部分由两个在结构上完全对称的多粒度特征提取模块组成，第一个通道的输入图像全部为可见光图像，而第二个通道的输入图像全部为红外图像。特征提取模块是一个卷积神经网络，它们的初始参数相同，但是权值不共享，因此在经过训练后，每个特征提取模块可以专注于提取可见光图像或红外图像的特征，每个通道输出 8 个 256 维的特征向量。然后，对于两个特征提取模块输出的两个模态行人的特征，我们使用一个共用的全连接层 FC-1 来提取可见光模态和红外模态的共性特征，对 FC-1 的输出进行 L2 正则化，正则化的目的是防止过拟合，正则化后的 8 个向量拼接成一个 2048 维的向量，该向量即图片的最后特征。为了训练该模型使得最终特征具有表征性，我们级联了异质中心损失（HC Loss）^[29]和身份损失来训练模型。由于身份损失函数的输入维度是训练集中行人身份的数量，因此我们添加了 FC-2 全连接层来满足身份损失的要求，FC-2 的神经元数量设置为行人身份的数量。在下面的小节中，我们将具体介绍多粒度特征模块和损失函数模块。

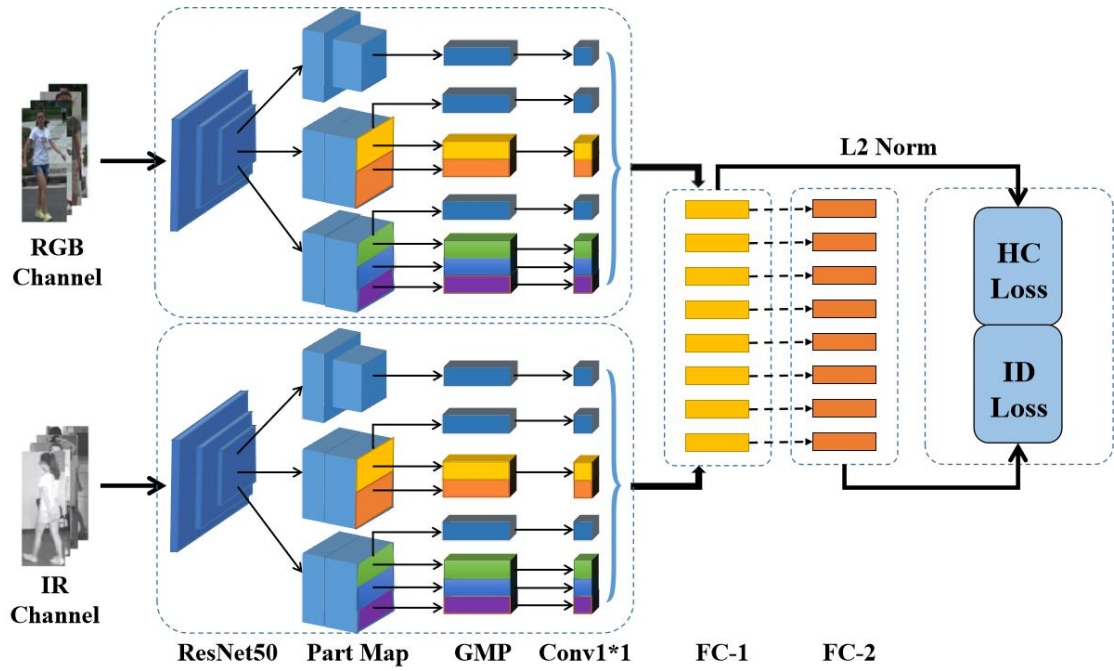


图 4-1 基于多粒度特征的双通道网络结构图

Figure 4-1. Structure diagram of dual-channel network based on multi-granularity features

4.1.1 多粒度特征模块

多粒度特征模块源于 Wang 等人^[17]提出的多粒度神经网络，该网络融合了行人的全局特征和局部特征，能够提取到更具判别性的行人特征，因此本章的特征提取模块选择使用多粒度神经网络的设计。

多粒度特征模块在 ResNet50 网络^[49]的基础上进行改进。首先，截取 ResNet50 网络中 res_conv4_2 以前的部分作为骨干网络。然后，一个全局特征分支和两个局部特征分支网络并行连接在骨干网络之后。全局特征分支连接 ResNet50 中 res_conv4 模块的剩余网络，而两个局部特征分支将 res_conv4 模块中最后一个卷积层的卷积步长由 2 改成 1，可以去掉这层卷积层的下采样过程，这样做的目的是使得到的特征图面积变为原来的四倍（长宽分别扩大一倍）。对比图 4-1 里“Part Map”模块上、中、下三个特征图，第一个分支是全局特征分支，输出的特征图尺寸缩小，而第二、三个局部特征分支的特征图尺寸不变。尺寸更大的特征图可以提取到更加细节的特征，也为后续的局部特征提供更多的分割空间。此外，对两个局部特征分支在“Part Map”模块后输出的特征图进行水平分割处理，分别从上到下分割为 2、3 个水平块。由于输入图片尺寸设置为 384×128 ，因此在经过卷积层、池化层的下采样操作后，“Part Map”模块后全局特征分支输出的特征图大小为 $2048 \times 12 \times 4$ ，两个局部特征分支输出的特征图大小为 $2048 \times 24 \times 8$ ，水平分割后的特征图大小分别为 $2048 \times 12 \times 8$ 、 $2048 \times 8 \times 8$ 。然后对全局特征分支和两个局部特征分支未经分割的特征图进行全局最大值池化（Global Max Pooling, GMP），得到 3 个 2048×1 的特

征向量，同时对局部特征分支分割后的五个特征图进行全局最大值池化，又得到 5 个 2048×1 维的特征向量。因此，在“GMP”层后输出为 8 个 2048 的特征向量。紧接着，使用一个 1×1 的卷积层对这些特征向量进行降维处理，该卷积层拥有 256 个卷积核，因此可以将 2048 维的特征向量降维至 256 维。 1×1 的卷积操作不仅可以降低特征向量维度，防止特征过于稀疏，还可以融合不同通道的信息、增加非线性特性。最终，多粒度特征模块将会把一张图像映射成 8 个 256 维的特征向量进行输出。

本章并联了两个多粒度特征模块，两个通道的多粒度特征模块分别输入可见光图像和红外图像，两个通道的结构一致、初始参数一致，但是参数不共享，也就是说，在经过训练之后，两个通道的参数将会有差异。第一个通道专门提取可见光图像的特征，而第二个通道专门提取红外图像的特征，因此两个多粒度特征模块的目的是为了提取两个模态的专属细节特征和共享特征。此外，两个全连接层“FC-1”、“FC-2”中的参数对于两个模态而言是共享的，目的是为了提取两个模态的共性特征。

4.1.2 损失函数模块

损失函数模块包含异质中心损失（HC Loss）和身份损失（ID Loss），我们称级联后的损失函数为 HCI Loss。身份损失函数可以有效的减小类内距离，在 3.1.3 节中已详细介绍，本节主要介绍异质中心损失函数。

每一个行人的不同图片在特征空间内存在一个类中心，由于跨模态行人重识别需要处理两个模态，因此每一个行人存在两个类内模态中心。异质中心损失函数的基本思想是以两个类内模态中心的平均距离为惩罚函数，最终拉近类内模态中心，降低异质模态造成的特征分散情况的发生，完成减小类内距离的目标。异质中心损失的公式为：

$$L_{HC} = \sum_{i=1}^U [\|c_{i,1} - c_{i,2}\|_2^2] = \sum_{i=1}^U \left[\left\| \frac{1}{M} \sum_{j=1}^M x_{i,1,j} - \frac{1}{N} \sum_{j=1}^N x_{i,2,j} \right\|_2^2 \right] \quad (4-1)$$

其中， U 表示训练批次中行人类别的个数， $c_{i,1}$ 和 $c_{i,2}$ 分别表示第 i 个行人在特征空间中的可见光模态类中心和红外模态类中心，而 $x_{i,1,j}$ 和 $x_{i,2,j}$ 分别代表第 i 个行人的第 j 张可见光图片和红外图片的特征向量，该特征向量是子特征向量经过 L_2 正则化后的结果，添加 L_2 正则化的目的是防止过拟合。

以加权求和的方式级联异质中心损失和身份损失：

$$L_{HCI} = \lambda L_{HC} + L_{id} \quad (4-2)$$

λ 是一个超参数。需要注意的是，这里的损失函数是对多粒度特征模块输出的 8 个向量的每一个进行计算。

4.2 实验与分析

4.2.1 实验设置

训练时，实验参数的设置尤为重要。数据输入方面，我们将输入图片的尺寸设置为 384×128 ，并且每次训练的批次选择策略与上一章的方法相同，我们将训练批次大小设置为 64，随机选取 4 个人，每个人 8 张红外图像和 8 张可见光图像。我们采用 SGD 优化器来更新网络中的参数，其动量设置为 0.9。学习率过大容易引起震荡，而太小又难以使网络收敛。因此我们设置初始学习率与网络层级有关，底层的初始学习率为 0.001，而最后一层全连接层的初始学习率设置为 0.01。此外，我们采用分阶段学习率衰减策略。每次试验我们一共训练 80 个周期，前 30 个周期训练时，学习率保持不变，30 到 60 个周期之间，学习率减小 10 倍，60 个周期后学习率再减小 10 倍。异质中心损失的加权参数 λ 设置为 0.5。

4.2.2 实验结果

我们在 SYSU-MM01 数据集上的实验结果如表 4-1 所示。我们在双通道多粒度（Dual-channel Multi-Granularity Network, D-MGN）的网络结构上，分别使用异质中心损失（HC）、身份损失（ID）、异质中心级联身份损失（HCI）进行对比。实验表明在级联损失函数的监督训练下，其性能远远超过单个异质中心损失函数或者身份损失函数。并且，我们发现在单独使用异质中心损失函数训练的过程中，损失值下降到 0.2 附近时开始震荡，因此判断模型没有收敛，导致精度非常低，我们猜想网络难以收敛的原因在于失去了身份损失的情况下，其行人特征分布在训练早期呈现无规律的零散状态，因此导致模态中心距离过大，引起网络的震荡，即使在经过长时间训练后，网络也无法达到最优状态。

此外，我们将 D-MGN 与单通道的多粒度网络（MGN）进行对比，我们在 MGN 上使用身份损失和上一章中的 HPI 损失进行训练，其结果展示在表 4-1 的前面两行。首先，我们发现 D-MGN 配合 HCI 损失的表现最佳。其次，D-MGN 与 MGN 在同一个身份损失的训练下，D-MGN 的精度要略胜一筹，在每个指标中都高出 MGN 一些，由此证明 D-MGN 与 HCI 损失搭配的有效性。同时我们发现单独使用 HC 损失训练 D-MGN 网络时精度非常差，而在加上身份损失之后才发挥作用，说明 HC 损失函数需要配合身份损失函数才能达到训练的效果。

我们也在 RegDB 数据集上进行了实验，该数据集的验证标准在 1.3.2 节中已经详细描述。实验结果如表 4-2 所示，与表 4-1 相比，我们增加了当前在 RegDB 数据集上表现优异的方法的对比，包括 zero-padding、eBDTR、D2RL。实验现象与 SYSU-MM01 上相似，在单独使用异质中心损失时，模型的实验结果变得很差，当级联异质中心损失和身份损失时，模型的性能得到了大幅度的提升。因此我们得出结论，在基于多粒度特征的双通道网络中，异质中心损失

需要搭配身份损失使用。目的是让身份损失大致的将不同类的特征在特征空间中区分开来，再利用异质中心损失缩短类内模态中心距离。

表 4-1 D-MGN 网络在 SYSU-MM01 数据集的实验结果

Table 4-1. Experimental results of D-MGN network on SYSU-MM01 dataset

方法	ASSS Rank1/mAP	ASMS Rank1/mAP	ISSS Rank1/mAP	ISMS Rank1/mAP
<i>MGN+ID</i>	27.29/27.60	31.05/21.65	33.47/43.08	38.19/32.80
<i>MGN+HPI</i>	39.77/41.12	44.86/34.88	44.06/54.52	50.55/44.90
<i>D-MGN+HC</i>	1.94/3.10	2.36/1.56	2.49/5.31	3.51/2.93
<i>D-MGN+ID</i>	28.32/31.82	32.66/35.13	34.76/46.65	39.19/35.92
<i>D-MGN+HCI</i>	43.49/45.23	47.72/49.21	46.36/56.43	51.05/57.49

表 4-2 D-MGN 网络在 RegDB 数据集的实验结果

Table 4-2. Experimental results of D-MGN network on RegDB dataset

方法	Rank1	Rank10	Rank20	mAP
<i>zero-padding</i>	17.8	34.2	44.4	18.9
<i>eBDTR</i>	34.62	58.96	68.72	33.46
<i>D2RL</i>	43.4	66.1	76.3	44.1
<i>MGN+ID</i>	26.2	61.8	72.5	36.9
<i>MGN+HPI</i>	46.8	70.2	81.1	40.4
<i>D-MGN+HC</i>	0.49	3.16	4.76	1.29
<i>D-MGN+ID</i>	28.45	51.36	61.94	31.91
<i>D-MGN+HCI</i>	48.93	75.29	83.64	47.38

4.2.3 排序结果可视化

为了更直观地了解模型在数据集中的排序结果，我们将训练完成的多粒度双通道网络（D-MGN+HCI）在 SYSU-MM01 的验证集上进行可视化。具体地，提取验证集里每一张行人的特征，计算查询目标库（query）中的红外图像与候选行人库（gallery）中每一张可见光图像的相似度，并按照从高到低的相似度进行排序。我们选取了几组图片进行展示，如图 4-2 所示，其中左侧一列表示待查询图片，右侧表示从 Rank1 到 Rank10 的排序结果。我们发现第一、二、四栏的行人表现良好，第三栏行人的最相似目标判断错误。然而，我们仔细观察第三栏的时候，发现 Rank1 的目标在肉眼上表现得也非常相似，他们的上衣服装颜色非常相近，红外图片中是白色的上衣，而实际上他穿的是黄色的上衣。由于红外相机和可见光相机成像原理的不同，红外图像会损失颜色和部分纹理

信息，这导致网络模型在识别时有一定的困难。而从前 Rank10 的结果来看，即使跨模态行人重识别存在种种困难，我们模型依然可以准确的找到正确身份的人物，这体现了该模型有可落地应用的潜质。



图 4-2 排序可视化

Figure 4-2. Ranking visualization

4.3 本章小结

本章对跨模态行人重识别的网络结构进行进一步研究，设计出基于多粒度特征的双通道网络（D-MGN）。实验证明双通道网络能够更好地提取特征，在 SYSU-MM01 数据集上取得了比上一章更好的表现，在 RegDB 数据集上也取得了优异的精度。我们分析了双通道网络的优势，使专属通道能专注于提取某个模态的特征，能够更好地发现不同模态图像的行人细节特征。此外，异质中心损失与身份损失搭配，能够取得惊人的效果。

然而，双通道网络也存在一些缺陷。由于前半部分的网络是参数不共享的，因此导致整个模型过于庞大，在计算资源受限的平台上运行会有些困难，并且在训练的时候特别占用服务器的显存。

第五章 基于梯度协调机制的身份损失函数的设计

5.1 引言

由于红外相机和热成像相机特殊的成像原理，它们拍摄的红外图像和热成像图像与可见光图像的差异非常大。因此，在第三、四章中，我们设计的模型和损失函数都有一个共同的目的：在特征空间中减小同类间的模态差异。然而在现实情况中，一些图像由于拍摄的原因在外观上与该行人的其他图像难以分辨，我们称之为难样本。大部分难样本会因为卷积神经网络强大的特征提取能力而被正确识别，而小部分难样本会被模型识别错误，因此我们猜想：在模型已经训练结束的情况下，在特征空间中仍存在一些难样本的特征脱离本类的聚集范围、分布在它类特征聚集处的情况，如图 5-1 所示，我们称这些脱离本类群体特征的特征点为“离群点”。

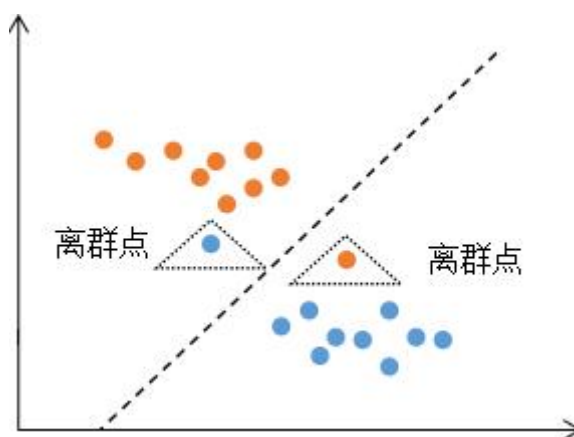


图 5-1 离群点特征分布示意图

Figure 5-1. Schematic diagram of outlier feature distribution

为了验证跨模态行人重识别中存在离群点的现象，我们进行一个特征可视化的实验。具体地，以 ResNet50 为基础网络，修改特征输出层的神经元个数为 2，分别在身份损失和 HPI 损失函数下训练。我们尝试了多次训练，并将行人特征在二维空间中进行可视化，统计发现存在离群点现象的模型占三分之一，而这仅仅是五个行人出现离群点的比例。存在离群点现象的特征分布如图 5-2 所示。

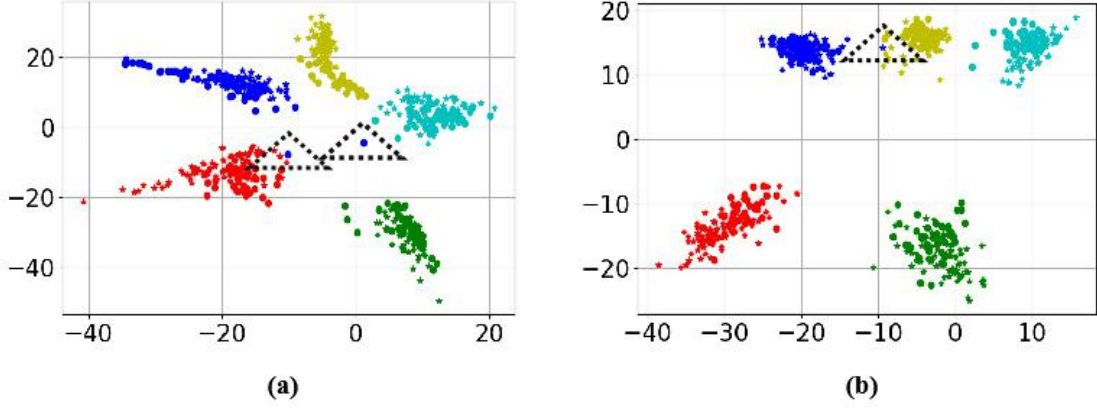


图 5-2 离群点特征分布真实图；(a) 身份损失；(b) 困难五元组级联身份损失

Figure 5-2. Real map of outlier feature distribution, (a) Identity Loss, (b) HPI Loss

离群点和易样本对模型的训练都是不利的。模型训练到后期阶段的时候，离群点的存在使损失函数的值不能下降，导致模型不能收敛，进而影响模型的性能。而易样本不需要在训练中被过多的关注，因为模型在经过短时间训练后，已经具有对易样本准确识别的能力。因此在训练后期，我们应该把训练的重心放在除易样本和离群点的训练样本上。

针对以上问题，本章将梯度协调机制^[50]结合到身份损失函数中，在训练过程中动态降低离群点和易样本在反向传播时的影响，从而加深模型的收敛程度。

5.2 梯度协调机制

梯度协调机制（Gradient Harmonizing Mechanism, GHM）发明于目标检测领域，通过动态调整难易样本的交叉熵损失值来协调反向传播过程中的梯度，以此降低目标检测任务中离群点和易分类点对训练造成影响，我们首次将其应用于跨模态行人重识别任务中。

首先来介绍在原论文中基于梯度协调机制的分类损失函数，该损失函数在基于 Sigmoid 的交叉熵损失函数上进行改进。二分类交叉熵损失函数定义为：

$$L_{CE} = \begin{cases} -\log(p), & \text{if } p^* = 1 \\ -\log(1-p), & \text{if } p^* = 0 \end{cases} \quad (5-1)$$

其中，一张图像经过卷积神经网络映射后在该类的神经元中输出 x ， p^* 表示图像 x 是否被模型正确识别，其取值为 1 或者 0。 p 表示模型的最大输出概率，最大输出概率 $p = \text{sigmoid}(x)$ 。在反向传播算法中 $\frac{\partial p}{\partial x} = p(1-p)$ ，根据链式法则得到交叉熵的梯度为：

$$\frac{\partial L_{CE}}{\partial x} = \frac{\partial L_{CE}}{\partial p} \cdot \frac{\partial p}{\partial x} \quad (5-2)$$

对对数函数求导，并其代入式(5-2)可得：

$$\frac{\partial L_{CE}}{\partial x} = \begin{cases} p - 1, & \text{if } p^* = 1 \\ p, & \text{if } p^* = 0 \end{cases} \quad (5-3)$$

在反向传播过程中的梯度如式(5-3)所示，由此定义梯度的模长为 g ，其公式为：

$$g = \left| \frac{\partial L_{CE}}{\partial x} \right| = |p - p^*| = \begin{cases} 1 - p, & \text{if } p^* = 1 \\ p, & \text{if } p^* = 0 \end{cases} \quad (5-4)$$

使用梯度模长 g 来量化样本的难易程度： g 越小样本越容易被训练，属于易样本；反之， g 越大训练难度越大，属于难样本；在训练后期梯度模长 g 仍然居高不下的样本属于离群点。

显然，离群点的梯度模长要比易样本的梯度模长大很多，因此无法直接将梯度模长按比例加权在损失函数上。为了在训练中同时减小对易样本和离群点的关注力度、把贡献梯度的任务交给对模型收敛有帮助的样本中，梯度协调机制应运而生。梯度协调机制的策略是优先衰减训练批次中数量更多的易样本或离群点，解决方案是将梯度模长均分为十个区间，以区间为单位定义梯度密度，最后在损失函数中除以梯度密度。梯度密度 $GD(g)$ 的公式为：

$$GD(g) = \frac{1}{l_{\omega}(g)} \sum_{k=1}^N \delta_{\omega}(g_k, g) \quad (5-5)$$

其中， N 表示训练批次中样本的个数， $\delta_{\omega}(g_k, g)$ 表示满足 $g \in (g - \frac{\omega}{2}, g + \frac{\omega}{2})$ 的梯度模长的个数，而 $l_{\omega}(g) = \omega$ ，即 $(g - \frac{\omega}{2}, g + \frac{\omega}{2})$ 区间的长度。因此， $GD(g)$ 表示梯度模长为 g 时的梯度密度。

最后，将交叉熵损失除以梯度密度，即基于梯度协调机制的分类损失 (L_{GHM-C})，其公式表达为：

$$L_{GHM-C} = \sum_{i=1}^N \frac{L_{CE}(p_i, p_i^*)}{GD(g_i)} \quad (5-6)$$

在训练后期，易样本和离群点的梯度密度要比其余样本高。因此，将损失值除以梯度密度，会减小易样本和离群点对损失值的贡献，从而使模型更多关注到有利于模型收敛的样本上。

为了将梯度协调机制应用在跨模态行人重识别任务中，来进一步提升训练效果，我们设计了一种基于梯度协调机制的身份损失：

$$L_{GHM-ID} = (1 - \rho)L_{ID} + \rho \cdot \lambda \cdot L_{GHM-C} \quad (5-7)$$

$$\rho = \frac{epoch}{max_epoch} \quad (5-8)$$

其中, L_{ID} 表示身份损失, ρ 代表训练的进度, L_{GHM-C} 表示基于 softmax 的梯度协调机制的分类损失函数, λ 表示权重。在训练早期, 身份损失占主要部分, 训练周期 $epoch$ 越大时, ρ 也越大, $1 - \rho$ 则减小, 意味着随着训练的推进, 身份损失的比重将减小, 而 L_{GHM-C} 的比重将增大。这样设计的原因是, 如果直接使用梯度协调机制的分类损失来训练, 网络将很难收敛。因为训练早期的时候, 模型不能辨别出难易样本, 需要用常规的身份损失进行辅助训练, 并且在训练的过程中逐渐升降二者的比重。

5.3 实验与分析

5.3.1 梯度模长可视化

为了探究难易样本的分布情况, 我们对样本的梯度模长进行统计并可视化。

首先, 我们以 ResNet50 为骨干网络, 搭配基于 softmax 的身份损失函数, 在 SYSU-MM01 数据集上进行训练。因为没有使用困难五元组损失函数等需要困难采样的损失函数, 我们将训练周期数 $epoch$ 调小, 设置为 10。初始学习率设置为 3×10^{-4} , 选择 Adam 为优化器, 批次大小设置为 64。

然后, 在训练结束后, 我们调用训练结果最好的模型作为验证模型, 在 SYSU-MM01 数据集的训练集上进行前向计算, 获取样本的输出概率, 从而计算该样本的梯度模长, 再将梯度模长进行统计, 数据集图像总量为 34167 张。由于样本的概率输出经过了 softmax 函数的映射, 因此它的分布在 (0,1) 区间内, 则梯度模长也被限制在该区间内。在统计梯度模长的时候, 我们按照四舍五入到小数点后一位的标准来保存梯度模长的数值, 并且区分可见光模态和红外模态, 从而统计出区间内某个梯度模长的不同模态的样本数量。

最后, 我们将梯度模长分布情况绘制成图 5-3。可以看出, 在模型训练到后期的时候, 易样本 (0~0.1) 占了多数, 并且离群点 (0.9~1) 比分布在中间模长的样本数量要稍高一些, 从而验证了我们的猜测。此外, 梯度模长从小到大的过程中, 红外样本数与可见光样本数的比值在增大, 说明红外样本的离群点出现概率会更多一些, 因为红外图像由于损失了颜色、纹理等信息, 导致更难识别。

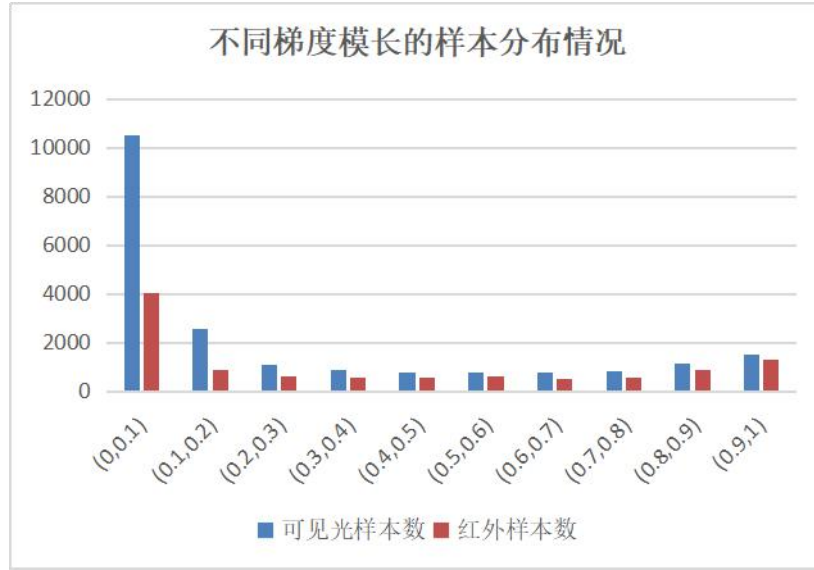


图 5-3 不同梯度模长的样本分布情况

Figure 5-3. Sample distribution of different gradient length

5.3.2 实验对比

为了验证基于梯度协调机制的身份损失在跨模态行人重识别领域中的有效性，我们进行了对照实验：使用 ResNet50 为骨干网络，分别辅助以 L_{GHM-C} 、 L_{ID} 、 L_{GHM-ID} 损失函数，在 SYSU-MM01 数据集上进行训练。除了损失函数不一样，其余环境设置一致。使用 Adam 优化器，学习率设置为 3×10^{-4} ，每个训练批次输入 64 张图片，最大训练周期为 30， L_{GHM-ID} 中的权重 λ 设置为 10，单独使用 L_{GHM-C} 时也需要乘以系数 10。

我们在 All-search Multi-shot (ASMS) 模式下进行比较，试验结果如表 5-1 所示。结果表明单独使用 L_{GHM-C} 损失函数进行训练时，网络不能收敛；使用 L_{GHM-ID} 损失函数进行训练时，相比于 L_{ID} 有了一定的提升。

表 5-1 梯度协调机制在 SYSU-MM01 上的实验结果

Table 5-1. Experimental results of the Gradient Harmonizing Mechanism on SYSU-MM01

方法	Rank1	Rank5	Rank10	mAP
ID	30.36	44.2	50.8	21.12
$GHM-C$	1.6	6.1	10.2	1.3
$GHM-ID$	32.16	46.08	52.16	21.52

此外，我们将训练过程中的 Rank1 绘制成折线图，如图 5-4 所示，横坐标代表训练周期 epoch，纵坐标分别表示 Rank1 精度。我们发现在训练的过程中，基于 L_{GHM-C} 的 ResNet50 模型精度一直很低， L_{GHM-ID} 和 L_{ID} 在训练中期有一定的起伏波动，但是在 epoch21 的位置都有一个较大的提升。

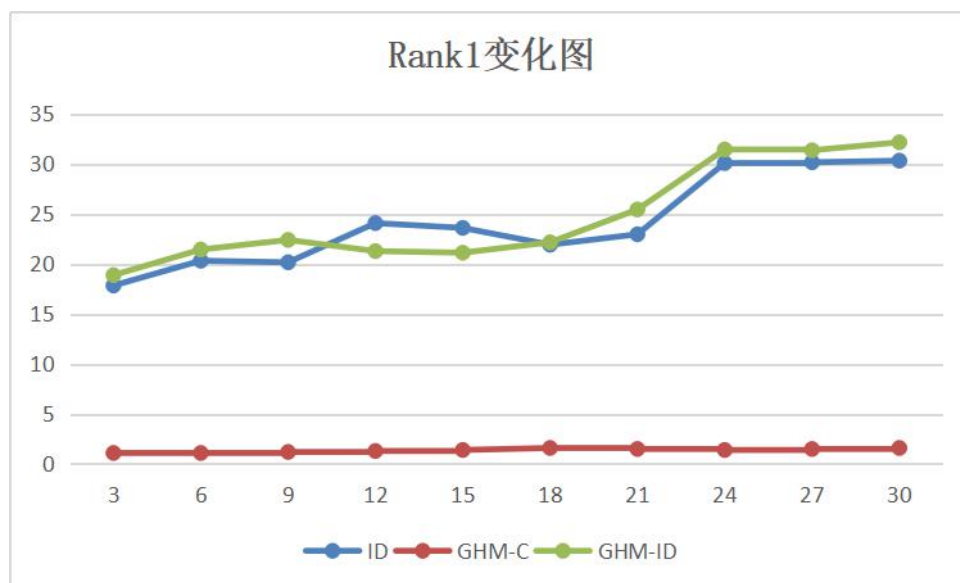


图 5-4 不同训练阶段的 Rank1 变化图

Figure 5-4. Changes in Rank1 at different training stages

5.4 本章小结

本章设计了基于梯度协调机制的身份损失函数，使模型在训练的过程中不过多的关注容易被识别的易样本，同时也不过多关注对模型训练提升不大的离群点。基于统计梯度密度的方式，针对性地对易样本和离群点的损失值进行削弱，实验表明在 ResNet50 为骨干的网络上有一定的提升。

第六章 总结与展望

6.1 总结

随着犯罪事件、弱势群体走失事件的频发，快速利用监控设备寻找到目标人物的需求亟需满足。行人重识别是基于监控图像或视频，利用计算机视觉寻找目标人物的方法。而如今丰富的监控设备种类造成监控图像具有模态多样化的特点，因此跨模态行人重识别具有重要的现实意义。常见的监控摄像头有可见光摄像头、红外摄像头和热成像摄像头，这些摄像头由于成像原理的不同，拍摄到的画面差异非常大。可见光图像具有丰富的颜色、纹理信息，拥有 RGB 三个通道的数据，然而可见光摄像头在黑暗的环境中无法拍摄到清晰的画面，而红外摄像头和热成像摄像头则可以在黑暗中工作。红外摄像头根据接受主动发射的红外光进行成像，因此红外图像没有颜色信息，并且失去了部分的纹理信息，以灰度图的形式储存图像。热成像摄像头根据物体散发的热红外成像，其画面容易受外界干扰，并且与红外图像一样，失去了颜色信息和部分纹理。因此，在跨模态的情境下匹配行人的身份是非常困难的。

此前，研究人员基于深度学习算法在跨模态行人重识别中取得了一定的成效，但是还不能达到非常准确的性能。本文基于深度学习设计了三种算法，在跨模态行人重识别数据集中取得了更高的准确度。本文的主要工作在以下三个方面：

(1) 设计了基于困难五元组的特征提取框架(HPILN)，该框架具有高精度、易复现、扩展性强的特点。该框架将现有的单模态行人重识别网络进行改进，迁移到跨模态行人重识别任务中作为特征提取模块。由于单模态行人重识别现有的网络针对行人特点进行了专门的设计，相比于分类网络能够更好的提取行人的特征。之后，针对提取好的特征向量，我们设计了一种困难五元组的采样方法，对训练批次中每张图像采样出一组困难五元组。接着，我们设计了一种新的损失函数：困难五元组损失函数。该损失函数加速了模型的收敛速度、提升了模型的精度。最后，我们使用身份损失函数级联困难五元组损失函数，实验证明进一步提升了精度，并且在该方法发表时达到了现有论文中的最高精度。

(2) 设计了基于多粒度特征的双通道网络。该双通道网络使用两个特征提取模块分别提取可见光图像和红外图像的行人特征，两个特征提取模块的结构相同，但是网络参数独立不共享，使特征提取模块能分别提取不同模态的细节特征。此外，我们的特征提取模块包含一个全局特征分支和两个局部特征分支，

融合全局特征和局部特征的方式可以提取到多粒度特征来保证提取到行人不同部位、不同尺度的细节特征。在损失函数方面，我们使用异质中心损失级联身份损失函数来训练模型。实验证明，该设计在跨模态行人重识别的公开数据集中达到了领先的精度。

(3) 研究了基于梯度协调机制的身份损失函数在跨模态行人重识别领域中的效果。我们分析了易样本和离群点在训练结束后的数量分布，得出了易样本和离群点比普通样本在单位梯度模长中比例更高、红外离群点比例高于可见光离群点比例的结论。为了使模型得到进一步的训练，我们将梯度协调机制融入身份损失函数，动态调整易样本和离群点在损失函数中的占比，从而减少模型对易样本和离群点的关注。通过对比试验证明，梯度协调机制在跨模态行人重识别中有一定的提升。

6.2 展望

本文提出的方法相比于跨模态行人重识别中的其他方法有了一定的提升，但是仍然存在可以进一步提升和完善的地方：

(1) 缺乏大规模的数据集。目前已公开的跨模态行人重识别 SYSU-MM01 和 RegDB 数据集的规模之和为三十万张图像，而且可见光图像的比例远高于红外图像。数据集存在样本少、不均匀的情况，导致训练的模型在应用落地的方面还存在一定困难。在后续的工作中，考虑使用数据扩充的方法，使用对抗生成网络将已有的单模态行人重识别数据集生成跨模态行人重识别数据集也是一个可以尝试的方向。如果硬件设备达到了要求，也可以尝试去采集真实环境中的跨模态行人重识别数据集。

(2) 本文设计的网络没有考虑移动端部署的问题，现有的一些监控设备已具备深度学习计算的能力，而在这些计算受限的设备上运行我们的模型会受到模型体积、参数数量的限制。因此，未来我们尝试将模型进行轻量化处理，以便于移动端的部署。

(3) 近期有学者提出素描图像与监控图像的跨模态行人重识别，该模式有一定的应用价值，比如刑侦时嫌疑人画面可能只来自目击证人绘制的素描图像。我们正在向相关研究人员申请该数据集，尝试将我们的方法应用在该领域中，并提出进一步的改进方案。

参考文献

- [1] 胡瑞敏. 大数据智慧城市与智能安防——行人重识别[J]. 智能建筑, 2014, (5): 44-47.
- [2] 罗浩. 深度学习时代的行人重识别技术[J]. 人工智能, 2019, (2019 年 02): 40-49.
- [3] Zheng L, Shen L, Tian L. Scalable person re-identification: A benchmark[C]. //Proceedings of the IEEE International Conference on Computer Vision. Piscataway, NJ: IEEE, 2015: 1116-1124.
- [4] Wei L, Zhang S, Gao W. Person transfer gan to bridge domain gap for person re-identification[C]. //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 79-88.
- [5] Zhao H, Tian M, Sun S. 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. Piscataway, NJ: IEEE, 2017: 1077-1085.
- [6] Wu A, Zheng W-S, Yu H-X. RGB-infrared cross-modality person re-identification[C]. //Proceedings of the IEEE International Conference on Computer Vision. Piscataway, NJ: IEEE, 2017: 5380-5389.
- [7] Swain M J, Ballard D H. Color indexing[J]. International journal of computer vision, 1991, 7(1): 11-32.
- [8] 张华. 基于空间颜色特征的行人重识别方法[J]. 华中科技大学学报: 自然科学版, 2013, 41(2): 209-212.
- [9] Lowe D G. Distinctive image features from scale-invariant keypoints[J]. International journal of computer vision, 2004, 60(2): 91-110.
- [10] Farenzena M, Bazzani L, Perina A. Person re-identification by symmetry-driven accumulation of local features[C]. //2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2010: 2360-2367.
- [11] 彭志勇, 常发亮, 刘洪彬. 基于 HSV 模型和特征点匹配的行人重识别算法[J]. 光子·激光, 2015, 26(8): 1575-1582.
- [12] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[C]. //Advances in Neural Information Processing Systems. New York, NY: Curran Associates, 2012: 1097-1105.
- [13] 陈兵, 查宇飞, 李运强. 基于卷积神经网络判别特征学习的行人重识别[J]. 光学学报, 2018, 38(7): 255-261.
- [14] Yi D, Lei Z, Liao S. Deep metric learning for person re-identification[C]. //2014 22nd International Conference on Pattern Recognition. Piscataway, NJ: IEEE, 2014: 34-39.
- [15] Li W, Zhao R, Xiao T. Deepreid: Deep filter pairing neural network for person re-identification[C]. //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2014: 152-159.
- [16] Sun Y, Zheng L, Yang Y. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline)[C]. //Proceedings of the European Conference on Computer Vision (ECCV). Piscataway, NJ: IEEE, 2018: 480-496.

- [17] Wang G, Yuan Y, Chen X. Learning discriminative features with multiple granularities for person re-identification[C]. //Proceedings of the 26th ACM International Conference on Multimedia. New York, NY: ACM, 2018: 274-282.
- [18] Xiao T, Li H, Ouyang W. Learning deep feature representations with domain guided dropout for person re-identification[C]. //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2016: 1249-1258.
- [19] Chen H, Wang Y, Shi Y. Deep transfer learning for person re-identification[C]. //2018 IEEE Fourth International Conference on Multimedia Big Data (BigMM). Piscataway, NJ: IEEE, 2018: 1-5.
- [20] Varior R R, Haloi M, Wang G. Gated siamese convolutional neural network architecture for human re-identification[C]. //European Conference on Computer Vision. Cham: Springer, 2016: 791-808.
- [21] 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. Piscataway, NJ: IEEE, 2015: 815-823.
- [22] Hermans A, Beyer L, Leibe B. In defense of the triplet loss for person re-identification[EB/OL]. 2017-03-27.[2020-03-12]. <https://arxiv.xilesou.top/abs/1703.07737>.
- [23] 郑伟诗, 吴岸聪. 非对称行人重识别: 跨摄像机持续行人追踪[J]. 中国科学: 信息科学, 2018, 48(05): 65-83.
- [24] Lin L, Wang G, Zuo W. Cross-domain visual matching via generalized similarity measure and feature learning[J]. IEEE transactions on pattern analysis and machine intelligence, 2016, 39(6): 1089-1102.
- [25] Ye M, Wang Z, Lan X. Visible thermal person re-identification via dual-constrained top-ranking[C]. //IJCAI. Menlo Park, CA: AAAI, 2018: 2.
- [26] Ye M, Lan X, Wang Z. Bi-directional center-constrained top-ranking for visible thermal person re-identification[J]. IEEE Transactions on Information Forensics and Security, 2019, 15: 407-419.
- [27] Feng Z, Lai J, Xie X. Learning modality-specific representations for visible-infrared person re-identification[J]. IEEE Transactions on Image Processing, 2019, 29: 579-590.
- [28] Xiang X, Lv N, Yu Z. Cross-modality person re-identification based on dual-path multi-branch network[J]. IEEE Sensors Journal, 2019, 19(23): 11706-11713.
- [29] Zhu Y, Yang Z, Wang L. Hetero-center loss for cross-modality person re-identification[J]. Neurocomputing, 2019.
- [30] Ye M, Lan X, Li J. Hierarchical discriminative learning for visible thermal person re-identification[C]. //Thirty-Second AAAI Conference on Artificial Intelligence. Menlo Park, CA: AAAI, 2018.
- [31] Nguyen D T, Hong H G, Kim K W. Person recognition system based on a combination of body images from visible light and thermal cameras[J]. Sensors, 2017, 17(3): 605.
- [32] Wen Y, Zhang K, Li Z. A discriminative feature learning approach for deep face recognition[C]. //European Conference on Computer Vision. Cham: Springer, 2016: 499-515.
- [33] Goodfellow I, Pouget-Abadie J, Mirza M. Generative adversarial nets[C]. //Advances in Neural Information Processing Systems. New York, NY: Curran Associates, 2014: 2672-2680.
- [34] Dai P, Ji R, Wang H. Cross-Modality Person re-identification with generative adversarial training[C]. //IJCAI. Menlo Park, CA: AAAI, 2018: 2.

- [35] Wang G A, Zhang T, Cheng J. RGB-infrared cross-modality person re-identification via joint pixel and feature alignment[C]. //Proceedings of the IEEE International Conference on Computer Vision. Piscataway, NJ: IEEE, 2019: 3623-3632.
- [36] Zhu J Y, Park T, Isola P. Unpaired image-to-image translation using cycle-consistent adversarial networks[C]. //Proceedings of the IEEE International Conference on Computer Vision. Piscataway, NJ: IEEE, 2017: 2223-2232.
- [37] Wang Z, Wang Z, Zheng Y. Learning to reduce dual-level discrepancy for infrared-visible person re-identification[C]. //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2019: 618-626.
- [38] Mcculloch W S, Pitts W. A logical calculus of the ideas immanent in nervous activity[J]. The bulletin of mathematical biophysics, 1943, 5(4): 115-133.
- [39] Cybenko G. Approximation by superpositions of a sigmoidal function[J]. Mathematics of control, signals and systems, 1989, 2(4): 303-314.
- [40] 周飞燕, 金林鹏, 董军. 卷积神经网络研究综述[J]. 计算机学报, 2017, 40(6): 1229-1251.
- [41] He K, Zhang X, Ren S. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification[C]. //Proceedings of the IEEE International Conference on Computer Vision. Piscataway, NJ: IEEE, 2015: 1026-1034.
- [42] Yu Q, Chang X, Song Y-Z. The devil is in the middle: Exploiting mid-level representations for cross-domain instance matching[EB/OL]. 2017-11-22.[2020-03-12]. <https://arxiv.xilesou.top/abs/1711.08106>.
- [43] Dai Z, Chen M, Zhu S. Batch feature erasing for person re-identification and beyond[EB/OL]. 2018.[2020-3-12]. <https://arxiv.org/abs/1811.07130>.
- [44] Chang X, Hospedales T M, Xiang T. Multi-level factorisation net for person re-identification[C]. //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 2109-2118.
- [45] Kingma D P, Ba J. Adam: A method for stochastic optimization[EB/OL]. 2014-12-22.[2020-03-12]. <https://arxiv.xilesou.top/abs/1412.6980>.
- [46] Hao Y, Wang N, Li J. HSME: hypersphere manifold embedding for visible thermal person re-identification[C]. //Proceedings of the AAAI Conference on Artificial Intelligence. Menlo Park, CA: AAAI, 2019: 8385-8392.
- [47] Kang J K, Hoang T M, Park K R. Person re-identification between visible and thermal camera images based on deep residual CNN using single input[J]. IEEE Access, 2019, 7: 57972-57984.
- [48] Selvaraju R R, Cogswell M, Das A. Grad-cam: Visual explanations from deep networks via gradient-based localization[C]. //Proceedings of the IEEE International Conference on Computer Vision. Piscataway, NJ: IEEE, 2017: 618-626.
- [49] He K, Zhang X, Ren S. Deep residual learning for image recognition[C]. //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2016: 770-778.
- [50] Li B, Liu Y, Wang X. Gradient harmonized single-stage detector[C]. //Proceedings of the AAAI Conference on Artificial Intelligence. Menlo Park, CA: AAAI, 2019: 8577-8584.

致 谢

三年研究生生活转瞬即逝，回顾这三年美好的学习时光，点点滴滴犹如昨日。在这三年里，不仅收获了知识，还感受到人与人之间美好的关怀和帮助。借这样的一次机会，向每一位帮助过我的老师、同学和家人，献上最诚挚的感谢！

首先，我要感谢我的导师赵云波教授。赵老师尊重学生的研究兴趣，不强制规定研究方向，这使我接触到了现在喜欢的研究领域。赵老师也给我们提供了良好的实验设备和学习环境，只要是对科研有帮助的地方，他都尽量满足我们的科研需求。在撰写学术论文的过程中，赵老师教会了我如何写科技文献、如何排版措辞、如何清楚的描述一个问题。赵老师花费大量的时间帮我修改论文，最终论文被录用，让我顺利达到毕业要求。

同时要感谢宣琦、陈晋音、俞山青老师。在宣老师团队学习的两年时光里，让我获益良多。宣老师是一名年轻的、对科研充满激情的老师，几乎每天都能在实验室看到他的身影。宣老师和陈老师在实验中给了我大量的指导，拓展了我的科研思路。团队里的老师和同学们对科研的热情也极大的感染了我，衷心祝愿 IVSN 团队越来越好。

其次要感谢我的室友们：徐彬彬、廖义辉、蒋晓奔。室友们为人随和，经常组织“搓一顿”的活动，在枯燥的学习时光中增添了不少乐趣。融洽的寝室氛围塑造了我的性格、收获了真挚的友谊。

此外，也要感谢我的同届：李灏、蒋传鹏、许德衡、潘晓康等，更要感谢我的师兄师姐和师弟师妹们，在科研过程中给了我很大的帮助。

最重要的是，感谢我的家人在二十年的学生生涯中对我的养育和支持，让我可以安心的学习。感谢家人对我付出的一切！

由衷地感谢评阅专家、老师和学者，在百忙之中审阅我的论文！

作者简介

1 作者简历

1995 年 6 月出生浙江省温州市。

2013 年 9 月——2017 年 6 月，绍兴文理学院机械与电气工程学院自动化专业学习，获得工学学士学位。

2017 年 9 月——2020 年 6 月，浙江工业大学信息工程学院控制科学与工程专业学习，攻读工学硕士学位。

2 攻读硕士学位期间发表的学术论文

[1] 李灏, 唐敏, 林建武, 赵云波. 基于改进困难三元组损失的跨模态行人重识别框架[J]. 计算机科学.2020.

[2] Yun-Bo Zhao, Jian-Wu Lin, Qi Xuan, Xugang Xi. HPILN: A feature learning framework for cross-modality person re-identification[J]. IET Image Processing, 2019, 13(14): 2897-2904 (SCI, IF=2.00)

3 参与的科研项目及获奖情况

4 发明专利

[1] 赵云波, 林建武, 李灏, 杨晨铭. 一种基于行人重识别的员工特定行为记录方法. 中国, CN109948490A[P]. 2019-06-28.

[2] 赵云波, 林建武, 李灏, 宣琦. 一种基于密集网络的多任务卷积神经网络顾客行为分析方法. 中国, CN109583942A [P]. 2019-04-05.

[3] 赵云波, 林建武, 李灏, 宣琦. 一种基于轻量化多任务卷积神经网络的导购行为分析方法. 中国, CN109544204A[P]. 2019-03-29.

[4] 赵云波, 林建武, 李灏. 一种基于人脸识别与行人重识别的特定目标跟踪方法. 中国. CN109344787A[P]. 2019-02-15.

[5] 赵云波, 林建武, 李灏, 宣琦. 一种基于 yolo 和多任务卷积神经网络的导购消极行为监控方法. 中国, CN109284733A[P]. 2019-01-29.

学位论文数据集

密 级*	中图分类号*	UDC*	论文资助
公开	TH741	681.2	
学位授予单位名称	学位授予单位代码	学位类型*	学位级别*
浙江工业大学	10337	工学硕士	全日制学术型硕士
论文题名*	基于深度学习的跨模态行人重识别研究		
关键词*	跨模态行人重识别, 跨模态图像检索, 智能安防, 度量学习		论文语种*
并列题名*	无		中文
作者姓名*	林建武	学 号*	2111703061
培养单位名称*	培养单位代码*	培养单位地址	邮政编码
浙江工业大学 信息工程学院	10337	杭州市潮王路 18 号	310032
学科专业*	研究方向*	学 制*	学位授予年*
控制科学与工程	计算机视觉	3 年	2020 年
论文提交日期*	2020 年 6 月		
导师姓名*	赵云波	职 称*	教授
评阅人	答辩委员会主席*	答辩委员会成员	
盲评	石崇源	宣琦, 赵云波, 陈晋音, 黄亮	
电子版论文提交格式: 文本 () 图像 () 视频 () 音频 () 多媒体 () 其他 ()			
电子版论文出版 (发布) 者	电子版论文出版 (发布) 地		版权声明
论文总页数*	51 页		
注: 共 33 项, 其中带*为必填数据, 为 22 项。			