

分类号: TP391

密 级: 公开

学校代码: 10363

学 号: 2220920108



安徽工程大学

Anhui Polytechnic University

硕士学位论文

题 目 遮挡场景下基于多粒度特征融
合的行人重识别

论 文 作 者 杨靖

指 导 教 师 谷灵康(副教授)

学 科 (专 业) 软件工程

研 究 方 向 领域软件工程

论文提交日期: 2025 年 5 月 9 日

分类号：TP391

单位代码：10363

密 级：公开

学 号：2220920108

题 目 遮挡场景下基于多粒度特征融合的行人重识别

英文并列

题 目 Person Re-Identification Based on Multi-Granularity
Feature Fusion in Occluded Scenarios

学生姓名： 杨靖

指导教师： 谷灵康(副教授)

专 业： 软件工程

研究方向： 领域软件工程

论文答辩日期：2025 年 5 月 10 日

安徽工程大学学位论文原创性声明

本人郑重声明：我恪守学术道德，崇尚严谨学风。所呈交的学位论文，是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已明确注明和引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的作品及成果的内容。论文为本人亲自撰写，我对所写的内容负责，并完全意识到本声明的法律后果由本人承担。

学位论文作者签名：杨靖

日期：2015 年 5 月 9 日

安徽工程大学学位论文版权使用授权书

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，允许论文被查阅或借阅。本人授权安徽工程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。

保密 ☐，在 ____ 年解密后适用本版权书。

本学位论文属于

不保密 ☒。

学位论文作者签名：杨靖

日期：2015 年 5 月 9 日

指导教师签名：王云康

日期：2015 年 5 月 9 日

遮挡场景下基于多粒度特征融合的行人重识别

摘 要

行人重识别也称行人再识别，作为智能监控系统的重要组成部分，在计算机视觉领域备受关注。旨在跟踪同一行人在多个监控摄像头中的身影。从计算机专业的角度来看，这一过程涉及到对行人图像或视频序列的特征提取、建模与匹配等一系列复杂操作。

目前，该技术已广泛应用于智能交通、流行病学调查、公共安全以及跨境追踪等领域。然而，由于行人在不同摄像头下的姿态、视角和遮挡等变化巨大，行人重识别任务面临诸多挑战，特别是对行人标注等因素的极度依赖。为了有效提升行人重识别任务的识别准确性，本研究基于深度学习这一强大的技术手段，深入探讨了多粒度特征融合的方法。深度学习凭借其强大的特征自动学习能力，能够从大量的行人数据中挖掘出深层次、更具代表性的特征。而多粒度特征融合方法则是通过整合不同尺度、不同层次的行人特征，充分利用图像中丰富的信息，旨在全面改善行人重识别算法的性能，以更好地应对复杂多变的实际应用场景。具体研究内容如下：

(1)遮挡会使行人的身体完全或部分被覆盖，这会给网络模型的行人身份识别带来挑战。目前的方法往往缺乏对遮挡区域的精细处理和识别，难以有效处理遮挡区域，并且对于部分遮挡的行人，在特征融合方面存在不足，可能无法有效提取、融合可见和遮挡区域的特征信息，使得最终的特征表示不够全面或准确。本文针对以上问题，考

虑到行人重识别任务在遮挡场景下准确率低的问题,提出一种结合随机遮挡和对比裁剪的无监督行人重识别。首先,根据现实场景将输入图像进行多种图案的随机遮挡,并使用对比裁剪的方法在保证正样本对语义一致的前提下,生成差异更大的行人图像样本,缓解遮挡和复杂背景对网络的影响;之后在对比学习模型中设计了 SC-CAResNet 网络进行多粒度特征提取,使其更加关注图像重要区域;最后联合多种损失函数对网络进行训练。

(2)现实环境中光照的变化会导致行人图像中的颜色、亮度和对比度发生变化,过强或过暗的光照都可能掩盖或改变行人图像中的细节特征,如面部特征、服装纹理等,这会使得算法难以准确区分不同行人之间的细微差别,从而降低行人重识别的识别精度。传统的暗光图像增强处理方法有:直方图均衡化、对比度拉伸、伽马校正等方法。这些方法虽然在一定程度上能够改善暗光条件下图像的质量,但也存在比如:不稳健、局部增强效果差、色彩失真和细节丢失等缺点局限,对行人重识别精度提升有限。针对以上问题,本文使用了深度学习的方法,设计了一种基于暗光增强和细粒度特征提取的行人重识别模型。首先利用暗光增强模块对输入的行人图像进行图像增强,优化图像信息,之后将图像输入改进后的 Vision Transformer 模型,该模型引入条件位置编码(Conditional Positional Encoding, CPE)和拼图补丁(Jigsaw Patches Module, JPM)促使网络关注行人图像的细粒度特征信息,增强对行人特征的提取能力。最后设计出 DAD-T(DropKey And Dynamic data pruning in Transformer, DAD-T)模块减少模型复杂度,缓

解性能下降的问题。以此在性能无损的情况下提高训练效率，加快训练速度。

最后，对本文使用到的几种方法进行了全面系统的归纳总结，同时，指出了现阶段行人重识别研究领域依然需要面对的一系列挑战。针对这些棘手的问题挑战，不仅展开了深入的讨论，分析了问题产生的原因和影响，还基于现有的研究成果和技术发展趋势，对未来行人重识别的研究方向进行了展望，探讨了可能的解决方案和研究趋势。

关键词：行人重识别，遮挡，多粒度，注意力机制，暗光

Person Re-Identification Based On Multi-Granularity Feature Fusion In Occluded Scenarios

ABSTRACT

Person re-identification, also known as person re-id, as an important part of the intelligent monitoring system, has received extensive attention in the field of computer vision. Its aim is to track the figure of the same pedestrian across multiple surveillance cameras. From the perspective of computer science, this process involves a series of complex operations such as feature extraction, modeling, and matching of pedestrian images or video sequences.

Currently, this technology has been widely applied in fields such as intelligent transportation, epidemiological investigation, public safety, and cross-border tracking. However, due to significant variations in the posture, viewing angle, and occlusion of pedestrians under different cameras, the task of person re-identification faces many challenges, especially an extreme dependence on factors such as pedestrian annotation. In order to effectively improve the recognition accuracy of person re-identification tasks, this study, based on the powerful technical means of deep learning, deeply explores the method of multi-granularity feature fusion. With its

powerful ability of automatic feature learning, deep learning can mine deeper and more representative features from a large amount of pedestrian data. The multi-granularity feature fusion method, by integrating pedestrian features at different scales and levels, makes full use of the rich information in the images, aiming to comprehensively improve the performance of person re-identification algorithms to better cope with complex and changeable practical application scenarios. The specific research content is as follows:

(1) In some cases, occlusion can completely or partially cover parts of a pedestrian's body, making it difficult for network models to identify the pedestrian's identity. Current methods often lack fine handling and recognition of occluded regions, making it challenging to effectively deal with occluded areas. Additionally, for pedestrians with partial occlusions, there may be shortcomings in feature fusion, which could result in ineffective extraction and fusion of visible and occluded region features, leading to incomplete or inaccurate final feature representations. To address these issues and considering the problem of low accuracy in person re-identification tasks under occlusion scenarios, this paper proposes an unsupervised person re-identification method that combines random occlusion and contrastive cropping. Firstly, various patterns of random occlusion are applied to input images based on real-world scenarios. Then, the method of contrastive cropping is utilized to generate pedestrian image

samples with greater differences while ensuring semantic consistency of positive samples, thereby alleviating the impact of occlusion and complex backgrounds on the network. Subsequently, an SC-CAResNet network is designed within the contrastive learning model for multi-granularity feature extraction, focusing more on important regions of the image. Finally, the network is trained using a combination of multiple loss functions.

(2) Changes in illumination in real environments can lead to variations in the color, brightness, and contrast of pedestrian images. Overly strong or dim lighting can obscure or alter detailed features in pedestrian images, such as facial characteristics and clothing textures. This makes it challenging for algorithms to accurately differentiate subtle differences between different pedestrians, thereby reducing the recognition accuracy of pedestrian re-identification. Traditional methods for enhancing dark images include histogram equalization, contrast stretching, and gamma correction. While these methods can improve the quality of images under low-light conditions to some extent, they also have drawbacks and limitations such as lack of robustness, poor local enhancement effects, color distortion, and detail loss, which limit the improvement of person re-identification accuracy. To address these issues, this paper employs a deep learning approach and designs a person re-identification model based on low-light enhancement and fine-grained feature extraction. Initially, an

enhancement module is used to optimize the information in input pedestrian images under low-light conditions. Subsequently, the improved Vision Transformer model is employed, incorporating Conditional Positional Encoding (CPE) and Jigsaw Patches Module (JPM) to focus the network on fine-grained feature information in pedestrian images, enhancing the capability to extract pedestrian features. Finally, the DAD-T (DropKey And Dynamic data pruning in Transformer) module is designed to reduce model complexity and alleviate performance degradation issues. This approach aims to improve training efficiency and speed without sacrificing performance, thereby enhancing the accuracy of person re-identification.

Finally, this paper comprehensively and systematically summarizes several methods used in the study. Meanwhile, it points out a series of challenges that the current research field of person re-identification still needs to confront. Regarding these tough challenges, this paper not only conducts in-depth discussions to analyze the causes and impacts of the problems, but also, based on existing research achievements and technological development trends, prospects the future research directions of person re-identification, and explores possible solutions and research trends.

KEY WORDS: Person re-identification, Occlusion, Multi-granularity, Attention mechanism, Low-light

目 录

安徽工程大学学位论文原创性声明.....	错误!未定义书签。
安徽工程大学学位论文版权使用授权书.....	错误!未定义书签。
摘 要.....	I
ABSTRACT.....	IV
目 录.....	VIII
第 1 章 绪论	10
1.1 研究背景及意义	10
1.2 国内外研究现状	11
1.2.1 基于有监督行人重识别.....	11
1.2.2 基于无监督域自适应的行人重识别.....	13
1.2.3 基于完全无监督的行人重识别.....	15
1.3 行人重识别的研究难点和挑战	16
1.4 本文研究内容及组织结构	18
1.4.1 本文研究内容.....	18
1.4.2 本文的框架结构.....	19
第 2 章 相关理论基础和技术	20
2.1 行人重识别基础知识	20
2.2 行人重识别常用数据集	20
2.3 行人重识别的评价指标	23
2.4 深度学习基础理论	24
2.4.1 深度卷积神经网络.....	24
2.4.2 ResNet 网络.....	30
2.4.3 ConvNeXt 网络.....	31
2.4.4 Transformer 相关概念及应用.....	31
2.4.5 注意力机制.....	33
2.5 对比学习	34
2.5.1 常用的对比学习框架.....	34
2.5.2 图像增强.....	36
第 3 章 结合随机遮挡和对比裁剪的无监督行人重识别	37
3.1 方法具体内容	38
3.1.1 方法概述.....	38
3.1.2 随机遮挡.....	39
3.1.3 对比学习.....	40
3.1.4 SC-CAResNet 网络模型.....	42
3.1.5 损失函数.....	45
3.1.6 消融实验.....	46
3.1.7 对比实验.....	47
3.2 本章小结	48

第 4 章 基于暗光增强和细粒度特征提取的行人重识别	49
4.1 方法具体内容	49
4.1.1 方法概述	49
4.1.2 改进的 ViT 行人重识别	50
4.1.3 暗光增强模块	51
4.1.4 条件位置编码	52
4.1.5 拼图补丁模块	54
4.1.6 轻量化模块	55
4.1.7 对比实验	57
4.1.8 消融实验	57
4.1.9 可视化热力图	58
4.2 本章小结	58
第 5 章 总结与展望	59
5.1 工作总结	59
5.2 未来展望	59
参考文献	61
攻读学位期间发表的学术成果	69
致谢	70

第 1 章 绪论

1.1 研究背景及意义

随着社会经济水平的提升和城市规模的不断扩大,日渐完善的天网监控系统成为智慧城市和平安中国建设的重要支撑。然而,社区、街道、广场等区域众多摄像头所产生的海量行人数据以及监控网络的复杂化,给人工处理数据带来了巨大挑战。这种情况不仅需要大量人力投入,而且耗费时间长,无法确保良好的识别准确度。

随着互联网技术的快速发展和大数据时代的到来,各种基于深度学习方法的行人重识别任务得到了广泛应用。作为图像检索任务的一个重要子领域,行人重识别可以弥补固定摄像头视野的局限性,并可以与行人检测和跟踪技术相结合^[1]。它已经成为构建“智慧城市”的重要组成部分,对于维护社会的和平与安定、促进公平与正义具有关键意义。这些基于深度学习方法的行人重识别准确率明显超越了传统方法,有效地减少了误判和漏检的情况。行人重识别技术整体实现过程如图 1-1 所示。

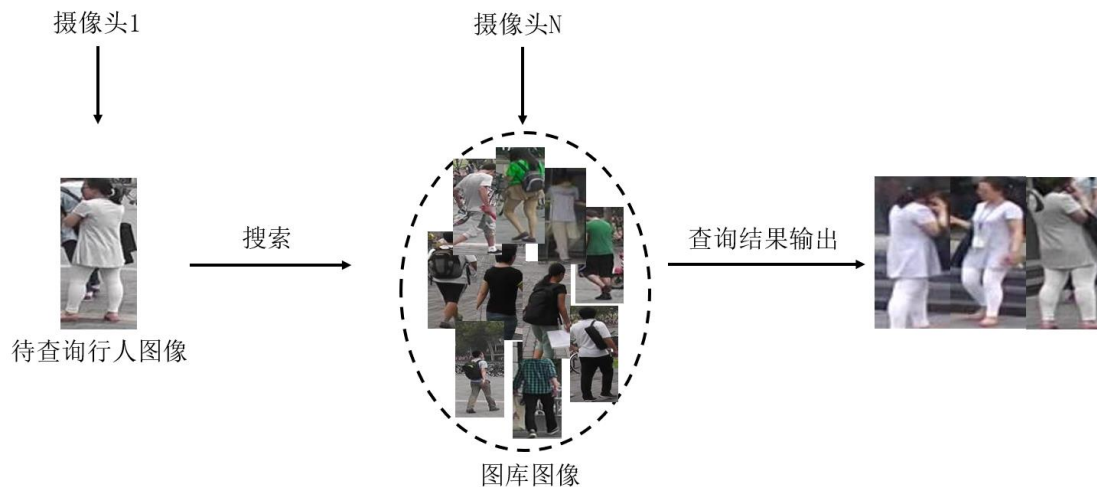


图 1-1 行人重识别任务实现过程图

Fig. 1-1 Person re-identification task implementation process diagram

在实际场景中,存在多个不同的监控摄像头,行人重识别通过计算机视觉技术利用这些摄像头采集的行人图像数据,旨在辨识目标行人是否出现在其中。

在当下的行人重识别研究与应用领域,有监督学习依旧是占据主导地位的方法。此方法依赖于标注完善的行人数据,这些数据详细记录了行人的各类特征与对应标签。借助这些标注数据,模型能够学习到不同行人之间的特征差异,从而完成训练过程,以具备在实际场景中准确识别行人的能力。然而,这些模型在扩展性和通用性方面存在一定的限制,并且标注行人数据集是一项繁重的任务。

为了应对这一挑战，使行人重识别更贴近实际应用，越来越多的研究者利用未标注的数据进行基于无监督方法的行人重识别研究。以便于在新场景中快速部署，这有助于适应复杂多变的实际环境。因此，在未来的研究中，基于无监督学习的行人重识别将发挥日益重要的作用，为解决实际场景中的行人识别问题提供更为灵活和高效的解决方案。

1.2 国内外研究现状

行人重识别算法经历了两个主要发展阶段：传统阶段和深度学习阶段。在早期的传统阶段，行人重识别的特征提取方法主要依赖于传统机器学习中的相关技术，旨在提取行人的基本视觉特征。然而这些方法不仅复杂且效率低下，难以应对不同摄像头下复杂场景的行人重识别任务，导致检测效果不理想，未能得到大范围推广，没有在研究领域取得更多进展。随着深度学习技术的兴起，行人重识别领域的研究思路变得更加开阔，基于深度学习的行人重识别方法逐渐成为主流研究方向^[2]。

以数据进行驱动的学习方式已成为构建高性能模型的基石，尤其在深度学习领域。需要大量高质量数据作为支撑，这些数据通常以图像、文本、音频等形式存在。通过不断迭代优化，模型逐步提高特征提取能力，从而增强对未知数据的预测效果。

随着数据规模的扩大，模型能够捕捉更丰富的信息，但也对标签质量提出了更高要求。早期的深度学习主要采用有监督学习方法，这种训练方式具有明确的学习目标。然而，数据标注过程费时费力，训练出的模型往往泛化能力有限，难以适用于标注成本高昂的大规模数据集。为了应对这些挑战，学者们提出了无监督学习的训练方法，不再人为设定标签，而是通过设定标注规则使模型自行为数据标记，这种标签被称为伪标签。

1.2.1 基于有监督行人重识别

一般而言，有监督学习这类方法将行人重识别任务视为验证或分类问题。常用方法有：特征表示学习和距离度量学习等技术。

基于特征表示学习的行人重识别方法主要依赖卷积神经网络，通过在训练集上自动学习行人的特征表示，并非直接学习行人图像间的相似性。这种方法被视为行人重识别领域的传统研究手段。在 2015 年，Zheng 等^[3]提出了特征表示学习的基准网络，如图 1-2 所示。

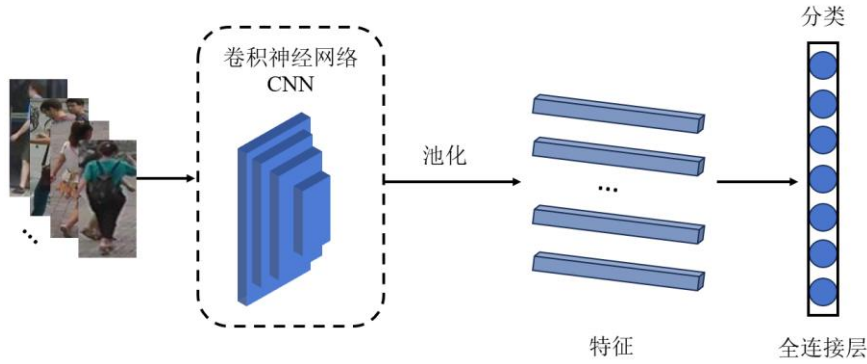


图 1-2 特征表示学习的基准网络图

Fig 1-2 Benchmark network diagram for feature representation learning

在计算机视觉领域，行人特征表示学习方法可细分为全局特征表示学习、局部特征表示学习以及序列特征表示方法。

局部特征表示学习方法利用人体姿态估计或粗略的水平分割技术来划分行人身体部位，示意图如图 1-3 所示。

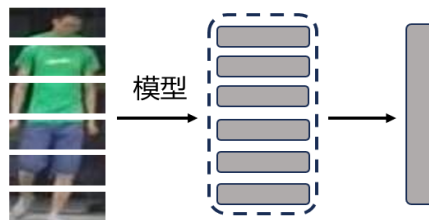


图 1-3 局部特征

Fig. 1-3 Local features

Sun 等^[4]提出了一种名为局部分块方法(PCB)，该模型首先提取行人图像的特征图，然后将这些特征图分割为不同的部分，并对每个部分进行局部池化操作。

Fu 等^[5]提出了水平金字塔匹配方法(HPM)，将每个行人图像视为多个水平金字塔块，并计算每个块的特征向量，以匹配不同图像中的区域，从而获得最佳匹配结果，将所有匹配结果合并，得出最终的行人重识别结果。

全局特征学习是通过卷积神经网络对整个图像进行特征提取，之后将其映射为一个低维特征向量的过程。这种方法有助于捕捉行人图像的整体信息，具有良好的识别效果和计算效率。如图 1-4 所示。



图 1-4 全局特征

Fig. 1-4 Global features

与局部特征学习相比，全局特征缺乏对图像局部细节的针对性处理，使得提取出的特征在描述图像时不够精细，从而造成其特征粒度相对较大。在面对遮挡或形变的情况时，其鲁棒性可能会受到影响。难以提取出极为细致、丰富的特征信息，尤其是对于存在遮挡或形变的行人对象。

在视频行人重识别数据集中，每个行人通过多个视频序列来表示，每个视频序列由多帧图像组成。

基于度量学习的行人重识别不同于早期依赖复杂设计的度量算法来衡量特征相似性的方法。这类方法的研究主要侧重于度量损失函数的构建，旨在缩小相似类别之间的距离，同时增大不同类别之间的间隔。常见的损失函数包括三元组损失^[6]、困难三元组损失^[7]、四元组损失^[8]以及对比损失^[9]等。

(1)三元组损失(Triplet Loss)

三元组损失如公式(1-1)所示：

$$L_t = (d_{a,p} - d_{a,n} + \alpha)_+ \quad (1-1)$$

该损失函数需要输入三张图片，分别为固定图片 a 、正样本图片 p 和负样本图片 n 。

(2)困难三元组损失(TriHard Loss)

困难三元组损失是由三元组损失改进而来的损失函数，它增加了区分难度。如公式(1-2)所示：

$$L_{th} = \frac{1}{P \times K} \sum_{a \in batch} \left(\max_{p \in A} d_{a,p} - \min_{n \in B} d_{a,n} + \alpha \right)_+ \quad (1-2)$$

(3)四元组损失(Quadruplet Loss)

四元组损失也是在三元组损失基础上改进得到的。通过比三元组损失函数多输入一张负样本图片，使得网络能够同时考虑正负样本的相对距离和它们之间的绝对距离。四元组损失如公式(1-3)所示：

$$L_q = (d_{a,p} - d_{a,n1} + \alpha)_+ + (d_{a,p} - d_{a,n2} + \beta)_+ \quad (1-3)$$

(4)对比损失(Contrastive Loss)

对比损失常用于孪生网络(Siamese Network)^[10]当中，可以很好地训练孪生网络。对比损失如公式(1-4)所示：

$$L_c = yd_{p1,p2}^2 + (1-y)(\alpha - d_{p1,p2})_+^2 \quad (1-4)$$

1.2.2 基于无监督域自适应的行人重识别

无监督学习和有监督学习之间的主要区别是：在训练过程中，是否有真实标签可供参考。无监督模型通过预设的标注规则自动为数据生成标签，这些标签通常被称为伪标签。无监督学习方法能够有效解决有监督学习训练中普遍存在的模

型泛化性能差和标注成本高等问题。常见的无监督学习方法有聚类 and 降维等，可辅助无监督标注。

无监督学习可根据是否使用带有真实标签的源域数据集将其分为无监督域适应和完全无监督两大类^[11]。

在无监督域自适应行人重识别任务中，为了提高模型的泛化能力，一般首先在源域数据上训练一个具有良好鲁棒性的行人重识别模型，随后通过微调该模型，以适应目标域数据。这种方法在迁移学习领域中扮演着重要角色，能够有效地解决源域和目标域之间数据分布差异的挑战，因此在深度学习领域逐渐成为一个关键研究领域。通过在源域学到的知识和特征，结合目标域的数据，实现模型的迁移和适应，为解决实际场景中的行人重识别问题提供了一种有效的方法。这种迁移学习方法的发展不仅推动了深度学习领域的进步，也为实际应用中的行人重识别任务带来了新的可能性和解决方案。图 1-5 为该类方法的示意图：

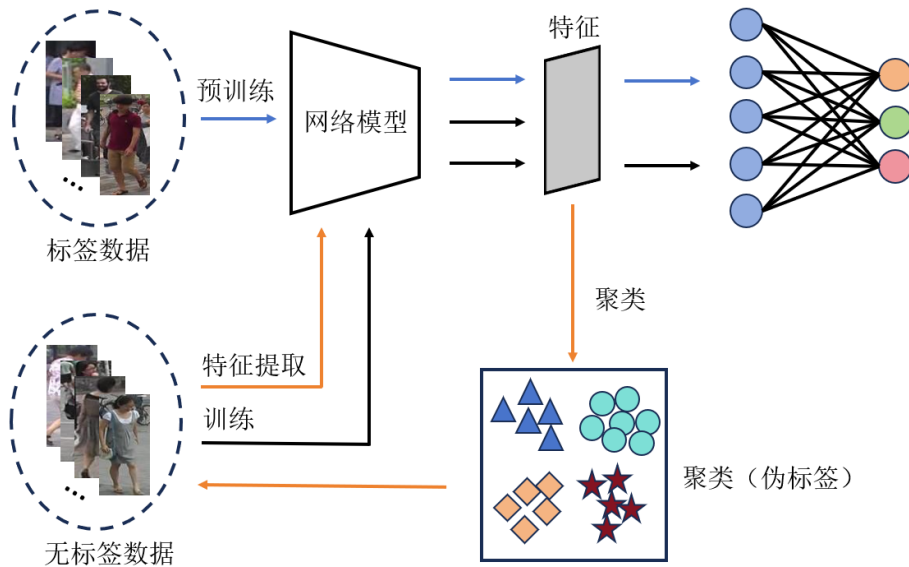


图 1-5 领域自适应行人重识别流程图

Fig. 1-5 Domain adaptive person re-identification flowchart

Yang 等^[12]为了通过挖掘全局和局部层面的视觉特征，探索未标记人员的潜在标签信息。提出了一种基于多标签学习的无监督框架。Wang 等^[13]提出了基于记忆的多标签分类损失(MMCL)，该方法能在统一框架中整合多标签分类和单标签分类。Zhai 等^[14]利用迭代密度聚类生成伪标签，有效地扩充了训练样本，从而增强网络模型的识别性能。Zou 等^[15]将行人图像特征空间划分为共享外观空间、源域结构空间和目标结构空间三个空间，并通过循环一致性图像生成任务，使网络能够自动区分特征，在共享外观空间执行对抗性对齐和自训练。Wang 等^[16]通过对相机感知信息的有效利用，借助基于相机感知的代理策略，在相机内和相机间的身份识别任务中取得更优的表现。Jia 等^[17]提出一种基于跨视角分布对齐(CV-DA)的无监督行人重识别方法，以减少无监督跨视角的影响。Lee 等^[18]引入

了一种基于摄像机标签的课程学习(CaCL)框架,利用人员图像的摄像机标签逐步将知识从源域传输到目标域。Chen 等^[19]将生成对抗网络 and 对比学习模块融合到一个联合训练框架中。生成对抗网络为对比学习提供在线数据增强,对比模块学习用于生成的视图不变特征。Ge 等^[20]提出了一种新颖的带混合记忆的自适应对比学习框架,有效提高了领域差距和聚类性能,充分利用所有有用信息。

1.2.3 基于完全无监督的行人重识别

虽然无监督域适应方法显著地提高了无监督行人重识别的性能,但这些方法的表现受源域数据集规模和质量的影响较为显著。为了解决这一问题,近年来一些研究人员开始关注如何利用未标记数据来训练高效的行人重识别模型。

完全无监督行人重识别直接将目标域内未标注的数据集用于模型训练。它充分挖掘目标域数据所蕴含的数据分布特性,从中学习具有区分能力的行人特征。这种方法的优势在于不依赖源域数据,直接利用目标域数据进行训练,有助于解决标记数据稀缺的问题,同时提高模型泛化能力和性能。通过完全无监督的训练方式,研究人员致力于探索如何在没有标签信息的情况下有效地学习行人特征,并取得令人满意的识别结果。这种方法不仅扩展了行人重识别领域的研究视野,也为面对实际应用挑战的解决提供了新的思路和可能性。未来,随着对无监督学习方法的进一步研究和优化,完全无监督行人重识别将在实践中发挥更为重要的作用,为智能监控系统和城市安全建设带来更大的帮助和推动。完全无监督行人重识别的示意图如图 1-6:

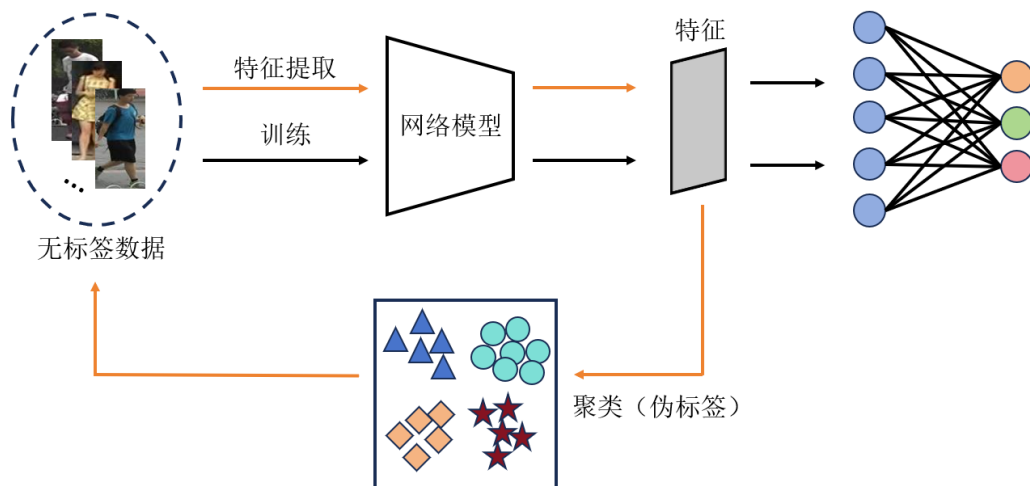


图 1-6 完全无监督学习行人重识别流程图

Fig. 1-6 Completely unsupervised learning person re-identification flow chart

针对聚类生成的标签难以确定的聚类个数和难以量化的聚类损失这两个问题,Zhang 等^[21]提出了双图对比学习(Dual-GCL),以充分利用实例之间的结构以及实例与聚类策略生成的质心之间的关系,有效解决了相机之间的域差异导致的类内方差较高的问题。Luo 等^[22]提出了用于无监督行人重识别的分层摄像头感知

对比扩展(HCACE),减少了聚类结果存在的噪声。Zeng 等^[23]提出的 HCT 方法将分层聚类与硬批次三元损失相结合,提高了现有方法中伪标签的质量。Yang 等^[24]提出了一种动态对称交叉熵损失(DSCE)来处理嘈杂样本,以及一种摄像头感知的元学习算法(MetaCam)来适应摄像头移位,提升了模型准确性,使得模型对摄像头变化敏感。Li 等^[25]在伪标签生成之前插入一个强大的特征校准模块,增加了伪标签生成和表示学习之间的联系。Xuan 等^[26]考虑摄像头之间的分布差异,提出了一种基于相机内相似度的伪标签生成方法,并将样本相似度计算分解为两个阶段:摄像机内部计算和跨摄像机间计算。Zuo 等^[27]考虑到大多数无监督行人重识别方法学习到的特征在嵌入空间中缺乏类内紧凑性和类间分离性,为了完整准确地表示簇中包含的信息并学习具有区分性的特征,提出了通过差异聚类代理的对比学习。Zhang 等^[28]提出了一种隐式样本扩展(ISE)方法来生成围绕簇边界的支持样本,有效避免了聚类有时会将不同的真实身份混合在一起的问题。Cho 等^[29]设计了一种新颖的基于部分的伪标签细化(PPLR)框架,该框架通过利用全局特征和部分特征之间的互补关系来减少标签噪声。Si 等^[30]充分考虑到考虑图像间和图像内的空间依赖关系,提出了一种新颖的空间驱动网络(SDN),用于从图像间和图像内的依赖关系中学习特别具有区分性的特征和丰富的语义信息。Si 等^[31]提出了一种新颖的混合对比模型(HCM),用于执行无监督行人重识别的身份级对比学习和图像级对比学习,充分探索了困难样本对之间的特征相似性。Wang 等^[32]提出了一个保持关系的特征嵌入(RPE)模型,利用样本之间的结构关系来提升无监督行人重识别方法的性能,无需任何样本注释。通过结合自动编码器和图自动编码器的优势,将样本内容和样本之间的邻域结构关系整合到特征嵌入的学习中。

1.3 行人重识别的研究难点和挑战

在行人重识别任务中,面临着许多困难和挑战,主要包括^[33]:

- (1)行人被遮挡。若遮挡区域较大或覆盖了关键判别特征,模型将难以提取有效信息,大幅降低识别精度。
- (2)图像分辨率低。摄像头的安装位置和设备质量对图像清晰度有重要影响。若摄像头距离目标较远或成像设备性能较低,获取的行人图像可能存在模糊现象,导致外观细节丢失。
- (3)光照环境的变化。不同环境下的光照会随着时间的变化而变化,室内室外的光线强度也具有差异。
- (4)视角变化。不同摄像头拍摄的行人外观可能发生显著变化,这种视角差异增加了特征匹配的难度。

(5)行人姿态变化。行人行进过程中，摆臂、迈步等动态动作会导致肢体位置变化，使摄像头捕获的行人外观存在显著差异。这些姿态变化会影响特征一致性。

(6)行人不对齐。部分数据集对框选中的部分可能只是行人的一个局部或者将行人周围大部分环境也包括在内。

(7)相似行人的干扰。不同行人在外观视觉上很相似的话就很难分辨是否是同一个行人。这些因素导致行人在匹配过程中难以对齐。

(8)数据集规模较小。深度学习模型需要大规模数据集支持，目前常用数据集规模有限，可能导致模型过拟合和泛化能力不足。当前研究者常用的数据集 Market-1501^[34]、DukeMTMC-reID^[35]数据量 3 万多张图像，CUHK03 数据集^[36]只有 1 万多张图像，即使目前较大的行人数据集 MSMT17^[37]，与 COCO 数据集^[38]相比数据量也太少，可能导致模型训练效果不佳。

通过针对这些问题的深入分析和解决方案的探索，可以不断提升行人重识别技术在实际应用中的效果和性能。

图 1-7 展示了几种行人重识别存在的问题。



图 1-7 行人重识别中存在的典型困难

Fig. 1-7 Typical difficulties in person re-identification

在无监督行人重识别中，除了上述提到的一般性挑战外，还会面临以下特定问题：

(1)缺乏真实身份标签是一个主要挑战。在没有真实身份标签的情况下，模型必须依赖于生成的伪身份标签进行训练，这会对特征表示学习产生困难和挑战。如果生成的伪标签不够准确或者质量不高，会直接影响到模型的训练和性能。

(2)噪声对伪标签的产生会有重要影响。特别是在行人重识别任务中,由于行人之间可能存在许多细微差异,从而使伪标签存在大量噪声。

(3)在行人重识别任务中,跨摄像机视角不变性特征表示的构建是模型性能优化的核心命题。其理论本质在于通过跨域特征空间对齐,提取对摄像机视角变换、光照条件差异等干扰因素保持鲁棒的身份判别性特征。然而,在无监督学习环境下,传统聚类算法(如 K-Means)生成的伪标签难以准确建模跨摄像机身份关联并且基于单摄像机聚类生成的伪标签存在跨摄像机误匹配风险。

(4)不同数据集间的差异。这个问题直接影响到模型的迁移能力。将在源域训练得到的模型应用于目标域时,模型的性能通常并不会理想,这主要是由于源域和目标域之间的差异导致了模型泛化能力的下降。

1.4 本文研究内容及组织结构

1.4.1 本文研究内容

本研究致力于增强行人重识别模型在复杂场景下的性能,特别关注遮挡和低光照场景。本文结合深度卷积神经网络(CNN)、注意力机制以及 Transformer 模型用以优化模型的表现,进一步提高行人重识别的精确度。下文将详细讨论研究的具体内容。

(1)为了减轻行人重识别任务对标记数据的依赖性以及缓解在遮挡场景下准确率下降的问题,提出了一种结合随机遮挡和对比裁剪的无监督行人重识别方案。该方法首先根据真实场景对输入图像进行多种随机遮挡,通过对比裁剪方法生成差异更大的行人图像样本,以减轻遮挡和复杂背景对网络的干扰。接下来,在对比学习模型中引入 SC-CAResNet 网络进行多尺度特征提取,使其更加聚焦于图像的关键区域。最后,结合多种损失函数对网络进行训练。实验结果表明,这种方法具有更强的鲁棒性。该方法显著提高了模型在遮挡场景下的行人重识别性能,为解决实际应用中遇到的遮挡问题提供了有效的解决方案。这种结合随机遮挡和对比裁剪的无监督行人重识别方案为行人识别的进一步发展和改进提供了新的思路和方法。

(2)为应对光照变化和色差等现实环境因素对行人重识别准确性的影响,提出了一种结合暗光增强技术与细粒度特征提取的行人重识别模型。首先,通过暗光增强模块对输入的行人图像进行增强,以优化图像信息。随后,将经过改良的图像输入到 Vision Transformer 模型中,该模型引入了条件位置编码(CPE)和拼图补丁模块(JPM),以促使网络关注行人图像的细粒度特征信息,增强对行人特征的提取能力。最后,设计了 DAD-T(DropKey And Dynamic data pruning in Transformer)模块,以降低模型复杂度,缓解性能下降问题,从而在保持性能的同

时提高训练效率,加快训练速度。在 DukeMTMC-ReID、MSMT17 和 Market-1501, 三个行人重识别公开数据集上进行了实验验证, 并与主流网络模型进行了比较。这种基于暗光增强和细粒度特征提取的行人重识别模型在处理光照变化和色差等挑战时展现出显著的优势, 提高了模型的鲁棒性和准确率。通过引入新颖的模块和技术, 该方法在提高行人重识别精度的同时, 也注重了模型的训练效率和速度, 为行人重识别技术的进步和应用提供了有力支持。

1.4.2 本文的框架结构

本文包括五个章节, 具体内容安排如下:

第一章: 绪论。在这一章节, 着重剖析了行人重识别研究的时代背景, 深挖其在现实场景中的重要价值。对国内外的研究现状进行了系统梳理, 同时细致分析了当下研究进程中遭遇的挑战与存在的问题。结尾部分, 对全文的行文架构、内容编排进行了简要介绍, 有助于快速把握文章脉络。

第二章: 行人重识别的理论背景与关键技术。本章首先阐述了卷积神经网络的核心理论。随后, 讨论了基于深度学习的行人重识别技术, 具体包括表征学习与度量学习的相关方法。最后, 本章还介绍了领域中常用的数据集和评估指标, 为后续研究提供了充分的理论基础和实验参考。

第三章: 结合随机遮挡和对比裁剪的无监督行人重识别研究。这一章主要针对有监督和现实遮挡场景下的行人重识别存在的问题, 通过设计新的提取网络并采用无监督学习的方式来提高识别精度。通过实验验证了方法的有效性。

第四章: 基于暗光增强和细粒度特征提取的行人重识别研究。这一章旨在解决暗光场景下行人重识别的难点。提出一种结合暗光增强模块和优化 Transformer 网络模型的解决方案, 减轻暗光图像对网络的干扰, 让网络进行细粒度特征提取。最后在相关数据集上进行实验。

第五章: 总结与展望。此章节里, 全面梳理了文中所提方法, 归纳其核心要点与关键步骤, 深度剖析研究过程中暴露出的短板与局限。与此同时, 立足当下研究基础, 结合行业发展趋势, 对未来潜在的研究方向展开合理展望, 为后续探索提供参考思路。

第 2 章 相关理论基础和技术

本章旨在介绍和分析行人重识别领域涉及的概念、常用技术以及实验中采用的数据集和评价指标。通过对这些内容的探讨，读者将更好地理解行人重识别领域的关键概念和技术，以及评价模型性能的标准。

2.1 行人重识别基础知识

行人重识别系统往往需要经历 5 个步骤：

(1)数据采集：为了确保行人数据的多样性，行人数据通常是通过监控摄像机采集的，这些数据可以是视频或图片。为了增加数据的多样性，这些监控摄像机通常被设置在不同场景和位置，这样有助于训练出具有鲁棒性的模型。这种多样性的数据采集方法有助于提高模型在各种场景下的泛化能力。

(2)行人检测：在计算机视觉领域中，通过矩形框将摄像机捕捉到的完整图像中的行人从背景中分离出来是至关重要的。精确的行人检测结果对于后续的行人跟踪和重识别具有关键意义，为确保后续任务的准确性和有效性提供了必要的基础。

(3)训练数据标注：这一流程通常在有监督的行人重识别任务中使用，数据标注方法包括身份标注和辅助信息标注。这些方法对于确保数据集的准确性和可用性至关重要，为后续的模型训练提供了基础。

(4)模型训练：在计算机视觉领域，通过采集数据来训练一个具有出色泛化能力和鲁棒性的网络模型是行人重识别任务的核心。训练模型通常涉及特征提取、损失函数定义、以及模型优化等关键步骤，这些步骤共同构成了模型训练的基本流程。这一过程对于确保模型在不同场景下的准确性和性能至关重要。

(5)行人检索匹配：首先，通过模型提取待查询行人图像与图库中每张图像的特征向量。然后，计算查询图像与图库图像之间的相似度。接下来，依据计算得到的相似度值，对所有图库图像进行排序，按照与查询图像的相似度从高到低生成排名列表。

2.2 行人重识别常用数据集

常用的行人重识别数据集包括图像数据集和视频数据集，图像数据集包括：Market-1501、DukeMTMC-reID、MSMT17 和 CUHK03 等。以上图像数据集介绍见表 2-1。

表 2-1 实验常用数据集

Table 2-1 Common experimental datasets

数据集	发布年份	训练集 人数/图片数	查询集 人数/图片数	相机数	场景
Market-1501	2015	751/12936	750/3368	6	室外
DukeMTMC-reID	2017	702/16522	702/2228	8	室外
MSMT17	2018	1041/32621	3060/11659	15	内外
CUHK03	2014	767/7365	700/1400	10	室内

Market-1501 数据集发布于 2015 年，在行人重识别研究领域有着举足轻重的地位。包含来自清华大学校园内 6 个不同监控摄像头的 32668 张行人图像，涵盖了 1501 位不同身份的行人。为研究人员提供了广泛且具有代表性的样本，助力在复杂现实场景下开展行人重识别相关研究，有效推动了该领域算法验证与技术革新。

下图展示了 Market-1501 数据集中部分行人图像的示例。



图 2-1 Market-1501 数据集图像展示

Fig. 2-1 Market-1501 dataset image display

DukeMTMC-ReID 数据集发布于 2017 年，该数据集的采集工作颇为精细，借助分布于不同位置的 8 台摄像机，全方位地记录行人影像。共包括 1404 位行人的 36411 张图像。这些图像涵盖了多样的拍摄视角与场景，为行人重识别研究提供了丰富的素材。

下图展示了 DukeMTMC-ReID 数据集中部分行人图像的示例。



图 2-2 DukeMTMC-ReID 数据集图像展示

Fig. 2-2 DukeMTMC-ReID dataset image display

MSMT17 数据集发布于 2019 年，它规模庞大，内容丰富多样，包含了 4101 位不同身份的行人，共计 126441 张图像。这些图像采集自 15 台不同摄像机，全方位覆盖了复杂多变的拍摄场景。通过这些图像，研究人员能够深入探索不同光照条件、拍摄角度以及行人姿态下的特征识别难题。

下图展示了 MSMT17 数据集中部分行人图像的示例。



图 2-3 MSMT17 数据集图像展示

Fig. 2-3 MSMT17 dataset image display

CUHK03 数据集发布于 2014 年，该数据集的行人图像，涵盖了多样化的场景和角度，采集自香港中文大学校园(CUHK)内的 5 对不同视角的摄像头。该数据集共包含 1467 位不同身份的行人。在训练集中，包含 767 个行人，而在测试集中则包含 700 个行人。

下图展示了 CUHK03 数据集中部分行人图像的示例。



图 2-4 CUHK03 数据集图像展示

Fig. 2-4 CUHK03 dataset image display

除此之外，行人重识别还有一些视频数据集：

MARS^[39]是一个大规模多摄像头行人重识别数据集，包含超过 100 万个行人图像和 20000 个视频序列。该数据集是由 6 个摄像机捕获的，涵盖了各种不同的场景、光线条件和姿势变化。

DukeMTMC-VideoReID^[35]是一个大型行人视频数据集，从 DukeMTMC 数据集中衍生而来，由 8 个摄像机拍摄多个角度的行人，最大特色在于其大规模的场景区覆盖和多样化的行人检测记录。

2.3 行人重识别的评价指标

在行人重识别算法中，常见的评价指标有：

(1)平均精度(Average Precision, AP)是一种针对单个目标类别的性能评估指标，用于评估单个目标类别的性能优良程度。

(2)Rank-k(前 k 项搜索结果排序)表示在前 n 个最相似的候选中是否包含了正确的匹配。Rank-k 可以表示为：

$$Rank - k = \frac{1}{p} \sum_{k=1}^p K \quad (2-1)$$

其中， p 表示为图像数量， K 为相似度前 k 项中有与目标行人身份相同的概率。通常情况下，常用的是 Rank-1 准确率，即最相似的候选中是否包含了正确的匹配。

(3)平均精度均值(mean Average Precision, mAP)，用于多类别的平均精度的计算。

$$mAP = \frac{1}{M} \sum_{i=1}^M AP_M \quad (2-2)$$

其中 M 表示图库中行人身份总数。 mAP 能很好的表示模型对于每个行人的综合识别效果。当类别为 1 的时候, mAP 退化成 AP 。

2.4 深度学习基础理论

此研究中所开发的行人重识别模型以深度神经网络为核心。在这一章节中, 将主要简要介绍涉及本文研究内容的相关技术和理论基础。

2.4.1 深度卷积神经网络

神经网络最初的灵感源自生物学领域对动物神经元的研究。在生物神经元中, 信号通过神经元的树突输入, 然后从轴突输出, 这些输出又作为输入连接到其他神经元细胞的树突上。神经网络随后被引入计算机科学领域, 其中卷积神经网络(CNN)作为前馈神经网络在多个前沿领域成果斐然, 在计算机视觉范畴, 助力图像识别、目标检测任务精度飞升; 在自然语言处理场景中, 推动文本分类、情感分析迈向新高度。图 2-5 清晰呈现了卷积神经网络的架构, 与传统神经网络相比, CNN 另辟蹊径, 大幅削减参数数量, 有效地缓和了过拟合难题。它之所以能够在众多领域大放异彩, 关键在于其独特的局部感受野、权值共享以及池化特性, 这些特性相辅相成, 赋予了 CNN 强大的特征提取能力和良好的模型泛化能力。

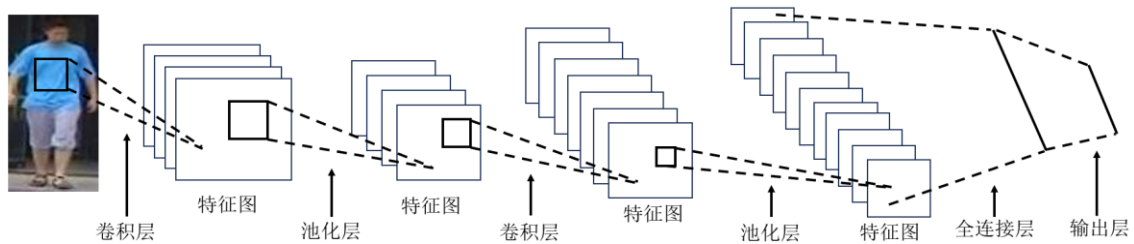


图 2-5 卷积神经网络框架图

Fig. 2-5 Convolutional neural network framework diagram

以下是对图中的卷积层、池化层、激活函数和全连接层的具体解释。

(一)卷积层是卷积神经网络的核心组成部分, 其主要功能是从输入的图像数据里精准地抽取出特征信息。卷积操作离不开卷积核, 在进行卷积运算时, 输入的图像会被转换为特征向量序列的形式。接着, 这个特征向量序列会与卷积核进行卷积计算, 经过一系列操作后, 生成输出特征映射。卷积核可以被视为一种特征提取器, 是一个小型矩阵, 其中的元素代表可学习的参数(权重), 用于对输入数据进行加权求和操作。在图像处理中, 卷积核也被称为滤波器(Filter)。通过卷积核与输入数据的局部区域进行卷积运算, 可以有效提取该区域的特征。图 2-6 展示了卷积层的运算过程:

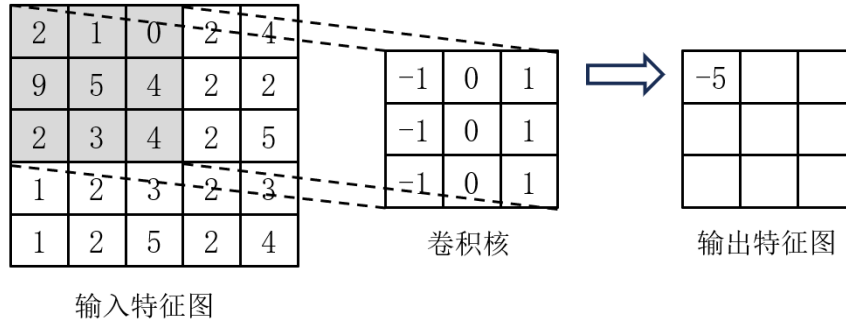


图 2-6 卷积层的运算过程图

Fig. 2-6 Graph of the convolutional layer operation process

输出特征图左上角计算方法为：

$$2 \times (-1) + 1 \times 0 + 0 \times 1 + 9 \times (-1) + 5 \times 0 + 4 \times 1 + 2 \times (-1) + 3 \times 0 + 4 \times 1 = -5$$

进行卷积操作可以设置以下参数：

数量：卷积层的输出通道数与所使用的卷积核数目紧密相关。每个卷积核会生成一个独立的特征图，输出通道数因此等于卷积核的数量。通过增加卷积核的数量，网络能够捕捉到输入数据中的多种特征，这有助于模型在不同层次上提取更加丰富的表示。

尺寸：在计算机视觉领域中，卷积核通常采用较小的尺寸，例如 3×3 、 5×5 等。这种选择有助于降低计算负荷，并且通过堆叠多个卷积层来获取更广泛的上下文信息。此外，小尺寸的卷积核也更容易学习到局部特征。

步长：在卷积过程中，步长代表了卷积核在处理输入特征图时从当前位置移动到下一个位置时沿宽度和高度方向各自移动的像素数。步长可以设定为不同的值，但通常情况下，宽度和高度方向的步长是相等的，以维持输出特征图的宽高比。

填充：填充是指在进行卷积操作之前，在输入特征图的周围添加额外的行和列，这些额外的行和列通常初始化为零值(或其他值)。这一步的目的在于控制输出特征图的尺寸，以防止在卷积过程中丢失边缘信息，并有助于在设计深层网络时保持输入和输出尺寸的一致性。

当输入图像大小是 $W \times W$ ，卷积核大小为 $F \times F$ ，步长为 S ，填充为 P ，最终的输出图大小为 $N \times N$ ，数学计算方法为：

$$N = \frac{W - F + 2P}{S} + 1 \quad (2-3)$$

(二)池化层：池化层在计算机视觉领域中也称为下采样层，其主要功能是通过减少特征图的尺寸来降低计算负担，从而提取出特征图的关键信息。通常，池化层紧随在卷积层之后，对卷积层的输出进行处理，不涉及任何需要学习的参数，而是对输入特征图中的像素值进行操作。池化层有助于降低输入特征的维度，这种方式可以有效地简化模型结构，提高模型的泛化性能。其具体操作基于局部

相关性的原理，通过对局部相关的一组元素进行采样或信息聚合，得到新的元素值。

池化层有多种类型，常见的包括最大池化、平均池化以及全局池化等。

最大池化操作是选取每个区域(即池化层的感受野)中的最大值作为输出。这种技术有助于保留输入数据中最为显著的特征，同时降低计算复杂度。

平均池化是通过取每个区域内的平均值作为输出。在图像数据中，噪声往往表现为一些随机出现的、与周围像素值差异较大的异常像素点。当采用平均池化时，这些噪声点的影响会被周围正常像素的数值所稀释。平均池化可以有效地减轻噪声点的突兀感，使得提取到的特征更加稳定和可靠。

全局池化则主要用于压缩特征图的空间信息，将其转换为固定长度的特征向量通常应用于分类任务的最后一层。全局池化分为对每个特征图的所有空间位置(高度和宽度)取平均值，得到一个标量的全局平均池化(Global Average Pooling, GAP)和对每个图的所有空间位置取最大值，得到一个标量的全局最大池化(Global Max Pooling, GMP)。前者能显著减少参数，防止过拟合，保留通道间的语义信息；后者则是保留最显著的特征，适用于需要突出局部关键信息的任务。

这些方法都能有效减少模型的复杂度，同时提升其在未知数据上的表现。研究人员可以根据具体的任务需求和数据特点，灵活选择合适的池化方式，优化卷积神经网络的性能，提升分类等任务的准确性和效率。

(三)激活函数：神经元接收多个输入信号，将其与权重相乘并求和后，通过激活函数生成输出信号。权重是神经网络中的关键参数，其数值大小决定了对应输入信号对神经元最终输出的影响程度。

激活函数对于卷积神经网络而言，具有无可替代的重要性，它赋予了卷积神经网络适应非线性模式的强大能力，堪称卷积神经网络架构中不可或缺的核心组成部分。倘若卷积神经网络仅仅是简单地堆叠多个卷积层，而没有引入其他非线性元素，那么整个网络模型的本质依然是线性的。从数学角度来看，无论堆叠多少层线性操作，其最终的输出结果依然可以通过一个线性函数来表示，这样的线性模型在面对现实世界中复杂多样的数据时，表现极为受限。

为了使网络更好地拟合复杂的非线性关系并提取深层次的特征，研究者通常会在多层卷积神经网络中引入非线性的激活函数。让整个多层卷积神经网络具备了强大的表达能力，能够高度准确地拟合现实世界中复杂的非线性关系，为实现高精度的行人重识别等计算机视觉任务奠定了坚实基础。

在图像处理中，常见的激活函数包括 Sigmoid、Tanh、ReLU、Leaky ReLU 等，如图 2-7 所示。

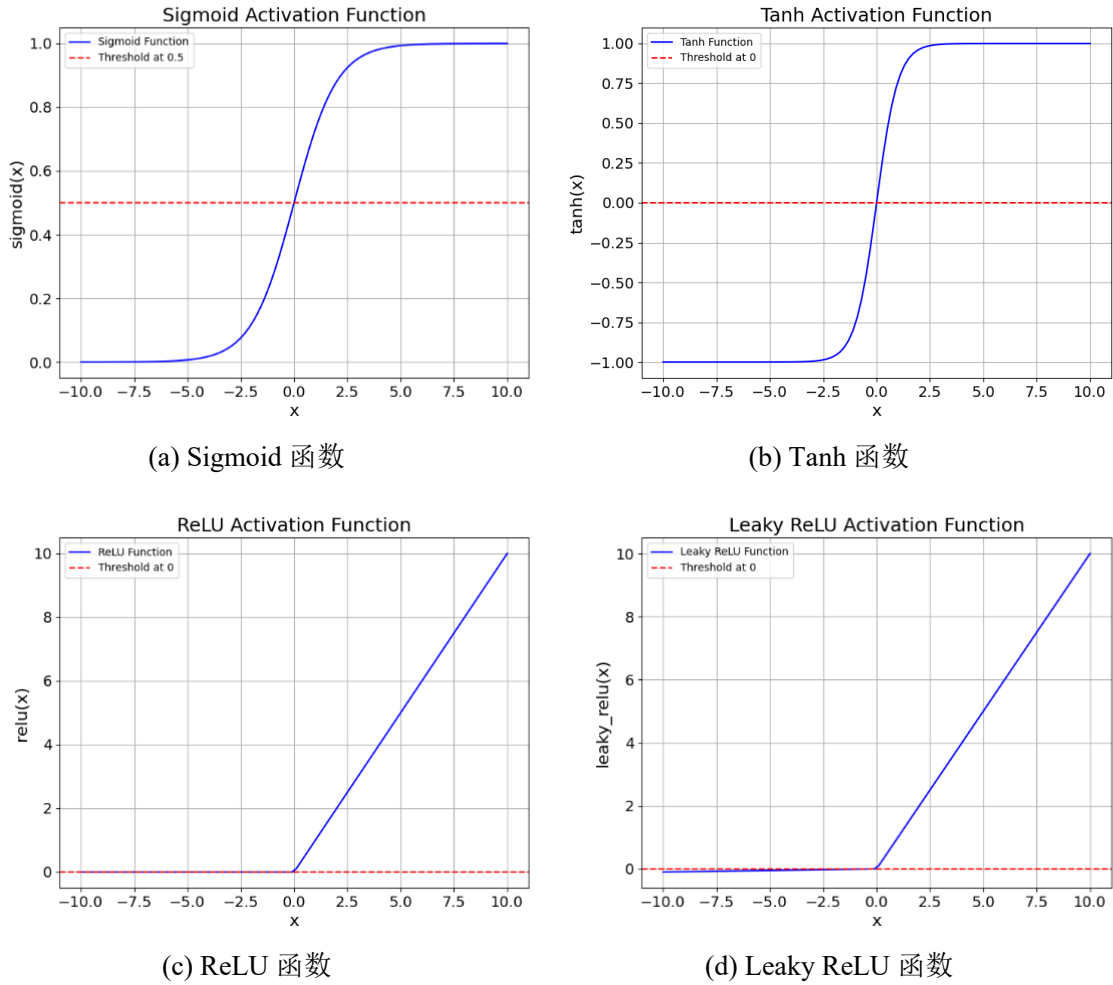


图 2-7 常见激活函数

Fig. 2-7 Common activation functions

Sigmoid 函数，又称为逻辑函数或 S 型函数，是神经网络中常见的激活函数之一，特别适用于二分类问题。但 Sigmoid 函数存在梯度消失问题，即在输入值非常大或非常小时，梯度接近于 0，导致网络在训练过程中难以有效更新权重。这一现象限制了网络的学习能力。梯度消失问题会影响网络的深度学习过程，导致梯度无法传播到较浅层，从而限制网络的表达能力。表达式为：

$$\text{sigmoid}(x) = \frac{1}{1 + e^{-x}} \quad (2-4)$$

其中 x 是输入值，函数输出值在 0 和 1 之间。

Sigmoid 函数具有以下特点：

(1)函数的曲线呈现 S 形状，具有单调递增性质。随着输入值的增大，输出值逐渐接近 1；随着输入值的减小，输出值逐渐接近 0。

(2)函数将实数输入映射到(0,1)的范围内，使其常被用于输出层，特别适用于二分类问题中，将输出解释为概率值。

(3)函数是一个光滑的函数，在整个定义域上具有可导性，使得在梯度下降等优化算法中应用更加方便。

(4)函数在输入值非常大或非常小时，梯度会接近于 0，导致梯度消失问题，限制了网络的学习能力和深度学习的表达能力。

(5)计算相对复杂，涉及指数运算，可能导致计算负担加重，尤其在深度神经网络中。

Tanh 函数，全称双曲正切函数(hyperbolic tangent function)，Tanh 函数将输入值映射到-1 和 1 之间，具有更好的对称性特点，这使得输出值的范围更广泛。相对于 Sigmoid 函数，Tanh 函数在训练过程中更容易实现收敛，这使得它在神经网络训练中具有一定的优势。

$$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \quad (2-5)$$

其中， e 代表自然对数的底， x 是输入值，其取值范围在-1 和 1 之间。

Tanh 函数具有以下特点：

(1)它是一种非线性函数，能够为神经网络引入非线性因素，从而帮助网络学习复杂的数据模式和特征

(2)函数的导数 e 在定义域内连续，这使得在反向传播算法中计算梯度更为方便，有利于神经网络的训练。

ReLU 函数，全称线性整流函数(Rectified Linear Unit)，是目前被广泛采用的激活函数之一。ReLU 函数在处理神经网络的过程中，对于输入值大于 0 的部分会保持原样输出，而对于输入值小于 0 的部分则输出 0。

$$\text{ReLU}(x) = \max(0, x) \quad (2-6)$$

其中， x 是输入值，当输入值 x 大于 0 时，ReLU 函数直接输出 x ；当输入值 x 小于或等于 0 时，ReLU 函数输出 0。

ReLU 函数具有以下特点：

(1)对于输入值小于等于 0 的部分，ReLU 函数输出为 0，具有稀疏性，有助于网络学习稀疏表示和特征选择。

(2)与 Sigmoid 和 Tanh 函数相比，ReLU 函数在反向传播过程中能够有效缓解梯度消失问题，从而使得深度神经网络的训练变得更加高效。

(3)由于 ReLU 函数的线性部分在正区间具有恒定梯度，因此可以加速网络的收敛速度，有助于提高训练效率。

Leaky ReLU 函数是对 ReLU 函数的一种改进，被称为带泄露线性整流函数(Leaky Rectified Linear Unit)，旨在解决 ReLU 函数可能导致的“死亡神经元”问题。当输入值小于 0 时，Leaky ReLU 函数允许一个很小的梯度值通过，从而避免“死亡神经元”问题。

Leaky ReLU 函数的数学表达式为：

$$\text{LeakyReLU}(x) = \begin{cases} x & , x > 0 \\ \alpha x & , x \leq 0 \end{cases} \quad (2-7)$$

其中， x 是输入值， α 是一个较小的常数，通常取值在 0 到 1 之间。当输入值 x 大于 0 时，Leaky ReLU 函数与 ReLU 函数一样，直接输出 x ；当输入值 x 小于或等于 0 时，Leaky ReLU 函数则输出 αx ，而不是像 ReLU 函数那样输出 0。

Leaky ReLU 函数广泛应用于深度学习和神经网络中的各个任务，包括图像处理、自然语言处理、强化学习和音频处理等。其优势主要体现在以下几个方面：

(1) Leaky ReLU 函数的设计旨在解决 ReLU 函数可能导致的“死亡神经元”问题。通过在负值区域允许一个小的斜率(leak)，Leaky ReLU 避免了神经元完全失活的情况。

(2) 与 ReLU 一样，Leaky ReLU 也是一种非线性激活函数，能够引入网络的非线性特性，有助于学习更复杂的数据模式。

(3) 计算简单且高效，只需要在负值区域应用一个小的斜率，使其在前向传播中易于实现。由于其简单性和解决“死亡神经元”问题的优势，Leaky ReLU 函数被广泛应用于深度学习模型中，并在一些情况下表现优于传统的 ReLU 函数。

(四)全连接层：在卷积神经网络的架构里，全连接层往往处在靠后的关键位置。它的核心任务是对先前各层提取出的特征执行加权和运算，进一步深度学习网络所捕捉到的特征信息，进而完成精准分类。在全连接层的独特构造下，相邻两层间的所有神经元都紧密相连，这种连接模式也就被称为 fully connected。图 2-8 为全连接层示意图，

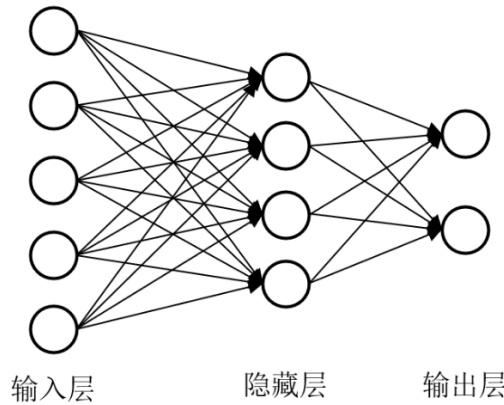


图 2-8 全连接层示意图

Fig. 2-8 Schematic diagram of the fully connected layer

全连接运算过程公式如下：

$$\text{output}(x) = f(W^T x + b) \quad (2-8)$$

式中， x 为输入向量， f 为激活函数， W^T 为权值矩阵， b 为偏置。

2.4.2 ResNet 网络

本小节重点是讲解深度学习行人重识别领域里应用最为广泛的骨干网络——ResNet (Deep Residual Network, ResNet)。ResNet 网络是 He 等^[40]于 2015 年提出的一种新的网络结构。在 2016 年, ResNet 网络首次被应用于行人重识别任务。Huang 等^[41]提出了一种基于局部对齐特征的方法, 用于提取行人图像的全局和局部特征。在行人重识别任务里, 大量的研究与实践都离不开 ResNet, 它凭借独特的网络结构和出色的性能, 成为众多模型搭建的基础架构, 在提取行人特征、提升识别准确率等方面发挥着关键作用。

然而, 在经典深度神经网络架构中, 梯度异常现象(梯度弥散/爆炸)严重制约了模型的深度扩展能力。针对这一瓶颈问题, ResNet 架构创新性地引入了残差学习机制, 使得网络能够更深。该网络通过跨层恒等映射(Identity Mapping)构建层级间短接通路, 使得前向传播过程中基函数可分解为恒等变换与残差分量叠加。这种结构在反向传播时形成梯度分流机制, 显著改善了深层网络中的信号退化问题。

实验表明, 残差结构特性使其在 ImageNet 等基准数据集上突破千层深度限制, 在行人重识别任务中, 基于 ResNet 骨干网络的特征提取器能够维持高维判别性表征能力, 显著提升跨摄像头目标的匹配精度。

残差学习模块的结构模型如下图所示, 简单有效地减轻了网络的退化问题。

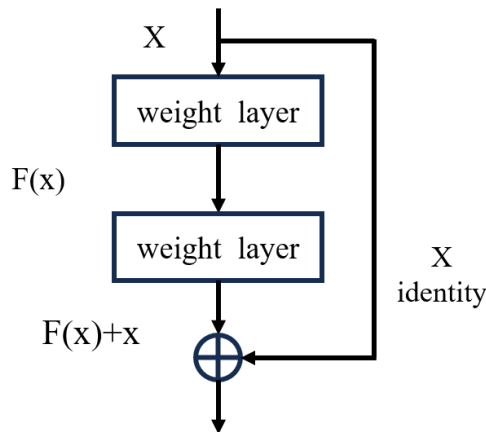


图 2-9 ResNet 网络残差模块

Fig. 2-9 ResNet network residual module

从 ResNet18 到 ResNet152, 模型的深度从 18 层增加到 152 层, 架构上也有所变化。

ResNet18 由于网络结构较浅适用于一般的图像分类任务, 特别是对于一些简单的图像分类问题或者计算资源有限的情况。

ResNet34 也适用于图像分类任务, 相比 ResNet18 更深, 有更多的层次。通常用于中等规模的图像分类数据集, 能够更好地捕捉图像特征。

ResNet50 在多个实际应用中得到广泛应用，可用于更大规模的图像分类数据集，如 ImageNet 等。

在需要更深的网络结构以处理更复杂的特征和模式时，通常会选择 ResNet101 和 ResNet152。

总的来说，ResNet 系列模型适用于图像分类任务，但具体选择哪个模型取决于任务的复杂度、数据集的规模以及计算资源的可用性。

2.4.3 ConvNeXt 网络

2022 年 Facebook AI Research 和 UC Berkeley 提出了 ConvNeXt^[42]纯卷积神经网络，ConvNeXt 在 ResNet50 模型的基础上，对标的是 2021 年非常出名的 Swin-Transformer^[43]，ConvNeXt 网络将 ResNet50 中的堆叠次数由(3,4,6,3)调整成(3,3,9,3)，并且将之前的卷积神经网络中最初的下采样模块修改成了一个卷积核大小为 4×4 ，步距为 4 的卷积层作为下采样模块。

对于 ConvNeXt 网络，作者提出了 T/S/B/L 四个版本，计算复杂度刚好和 Swin Transformer 中的 T/S/B/L 相似。在相同的 FLOPs(浮点运算次数)下，ConvNeXt 网络在推理速度与准确率方面展现出显著优势，以 ImageNet 22K 数据集为测试基准，ConvNeXt 达到了 87.8%的准确率，这一成果在图像识别领域具有重要的参考价值。

2.4.4 Transformer 相关概念及应用

2017 年，Vaswani 等^[44]提出的 Transformer 模型在自然语言处理领域取得了显著进展，成为了机器翻译任务中的主流模型。Transformer 框架由编码器和解码器两部分组成，框架流程图如图 2-10。

在模型接收信息后，编码器对信息进行初步编码，生成编码序列，随后解码器对序列进行解码，生成相应任务的解码序列。Transformer 架构的创新性体现在于编码器-解码器架构里融入了多头注意力机制。这一机制让模型能像人类一样，在处理信息时，自动聚焦于关键内容，显著提升信息处理效率。每个模块中，多头注意力机制率先捕捉关键信息，再通过归一化操作稳定数据分布，最后借助前馈神经网络对信息深度加工。在实际应用中，开发者能依据任务复杂度、数据量等情况，灵活增减模块数量，精准调控模型深度。

这种可调节的架构设计，让模型在层层递进的运算中，不断筛选、提炼，挖掘出数据里最具价值的信息，为各类复杂任务提供有力支持。

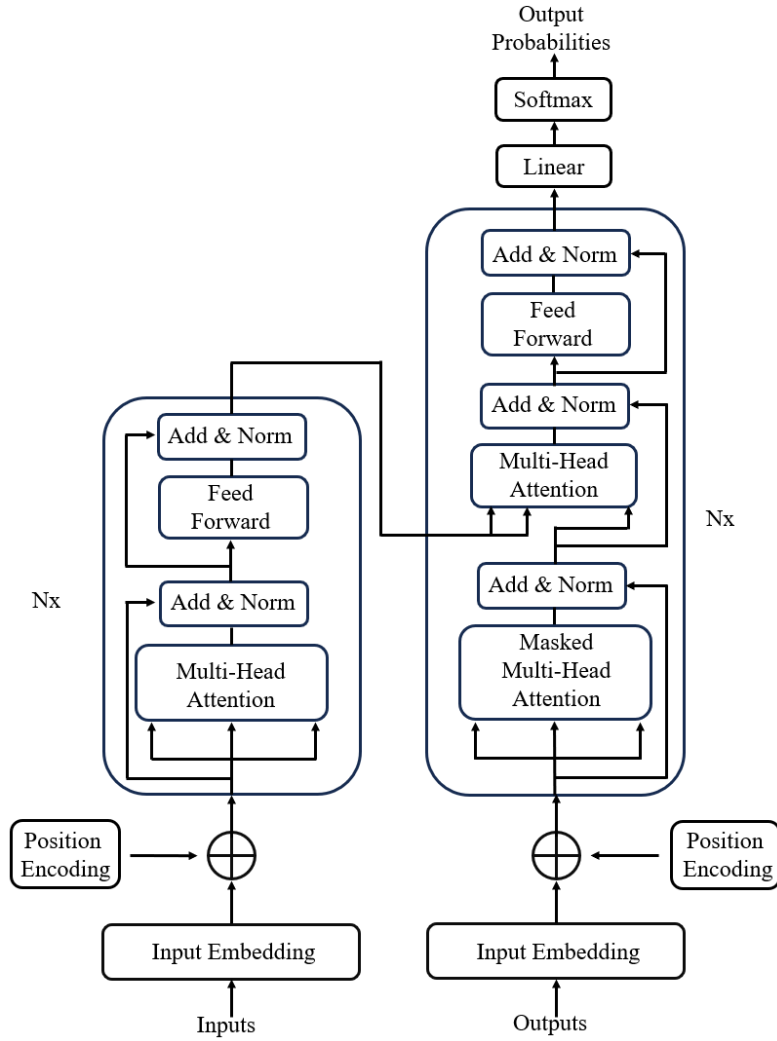


图 2-10 Transformer 模型框架

Fig. 2-10 ResNet network residual module

随着 Transformer 模型在计算机视觉领域的卓越性能和简洁结构逐渐被认可，越来越多的研究者开始探索 Transformer 在视觉任务中的应用和潜力。Dosovitskiy 等研究者提出了 ViT(Vision Transformer)模型^[45]，这是一种完全基于自注意力机制的网络结构，在大规模数据集上展示了不依赖 CNN 的 Transformer 在分类任务中表现良好的证据。ViT 首次将 NLP 领域的 Transformer 模型引入图像分类任务，表明 Transformer 模型不仅在自然语言处理领域有广泛应用，同时具有高度的可迁移性和通用性，为计算机视觉领域带来了新的思路和方法。

He 等^[46]提出了第一个基于 Transformer 的行人重识别框架，名为 TransReID，包含拼图补丁模块(JPM)和辅助信息嵌入(SIE)，有效提取了局部特征并将其聚合以获得具有全局上下文的鲁棒特征。Li 等^[47]构建了一个双分支网络架构，研究探讨了从纯 Transformer 网络中提取多粒度特征的方法，以解决无监督行人重识别问题，所提出的方法在无监督方法中明显优于最先进方法，大大缩小了与监督

方法的差距。王等^[48]提出一种融合卷积注意力和 Transformer(FCAT)架构的方法,以增强 Transformer 对局部细节信息的关注,所提方法在非遮挡数据集上取得了与现有方法相当的结果。赵等^[49]提出改进型 Transformer 编码器和特征融合的行人重识别算法,有效解决了 Transformer 编码器在行人重识别中因图像块信息丢失以及行人局部特征表达不充分导致模型识别准确率低的问题。Zhu 等^[50]首次在 Transformer 架构中引入了一种对齐方案,并提出了自动对齐 Auto-Aligned Transformer(AAformer)以便在补丁级别自动定位人体部位和非人体部位。

2.4.5 注意力机制

在计算机视觉体系中,注意力建模(Attention Modeling)已成为提升深度网络表征能力的关键技术路径。其核心思想源于生物视觉系统的选择性注意力机制——当人类观察包含行人的复杂场景时,视觉皮层通过视网膜中央凹动态聚焦于显著性区域(如人体轮廓、衣着纹理),同时抑制背景干扰信号的神经传递。

受此启发,这种机制被广泛应用于深度学习的各个领域,证明了其有效性。在图像领域的研究中,注意力机制通常通过增强特征的空间位置和通道维度来提高特征的表征能力。

该方法在行人重识别任务中展现出独特优势:通过动态特征调整强化身份敏感区域(如头部、鞋类等跨摄像头稳定特征)的响应强度,有效缓解姿态变化与背景混杂带来的域间差异。

由于图像中不同区域对于识别任务的贡献不同,背景区域的特征对行人重识别任务的贡献较小,甚至可能带来负面影响。空间注意力机制通过对输入特征进行加权,使得网络能够专注于图像中的重要区域,并减少对不相关区域的关注。这种机制依赖于学习不同空间位置之间的关系,进而对特征图的各个位置赋予不同的权重,从而提升关键区域的特征表达能力。通过引入专门的网络模块,空间注意力能够自动学习特征空间的分布模式,并生成空间注意力映射,帮助网络专注于最具信息量的区域。这种方法使得网络能够从加权后的输入特征中提取出更具判别力的特征,进而改善行人重识别任务的性能。

随着卷积神经网络的深度增加,特征图的通道数逐渐增多。每个通道代表一个特征提取器所捕捉到的特定特征,其中一些通道的特征对任务有用,而其他通道则可能对任务贡献较小。通道注意力机制的目标是通过为有助于任务的通道分配更高的权重,来增强这些特征的表示。它通过评估通道之间的关系,学习每个通道的重要性,并据此调整权重,从而使得网络能够更有效地聚焦于对任务有用的特征,同时抑制不相关或无关的通道信息。

在跨域 ReID 任务中,通道注意力机制展现出独特优势:

(1)光照不变特征增强:通过强化对光照鲁棒的通道(如 HSV 颜色空间中的色调通道),抑制亮度敏感通道的响应。

(2)跨模态特征对齐：在可见光-红外行人重识别中，采用双流通道注意力网络分别处理不同光谱输入，通过权重共享机制隐式对齐特征分布。

(3)细粒度鉴别学习：基于通道注意力生成部件敏感特征掩膜(如鞋类、背包区域)，结合局部特征匹配策略提升检索精度。

近年来，出现了很多结合注意力机制和行人重识别的相关研究，Tay 等^[51]提出了属性注意力网络(AANet)，将人员属性和属性注意力图整合到一个分类框架中，这是一种全新的分类框架。可以进行包括性别、头发长度和服装长度在内的人物属性预测。Song 等^[52]设计了一个掩模引导的对比关注模型(MGCAM)，从身体和背景区域分别学习特征，有效应对了复杂背景对识别精度的影响。Zhou 等^[53]提出了一个基于深度学习的新型人体注意力框架(AFPB)使神经网络不仅能够区分一个人的身份，还能确定样本来自遮挡数据分布还是完整身体数据分布。何等^[54]引入跨通道交互的注意力机制，增强行人特征的鲁棒提取能力，获得更具辨别力的特征表达。

2.5 对比学习

对比学习方法作为一种无监督学习方法，能够有效利用大规模无标签数据训练模型，更具实用价值。对比学习可被视作为字典查询任务训练编码器的方法^[55]，旨在拉近正样本的距离，同时将负样本推开。对比学习模型的流程图如下：

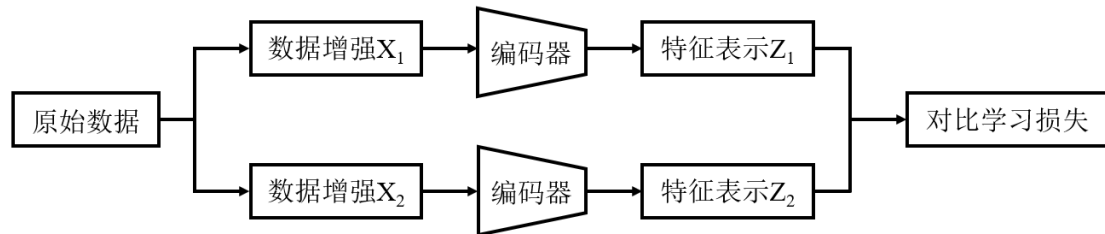


图 2-11 对比学习模型的流程图

Fig. 2-11 Flowchart of the contrastive learning model

通过观察模型，我们可以发现，原始数据在经历数据增强后，生成了增强数据 X_1 和 X_2 。这两组数据分别通过映射函数 f (编码器/模型)转化为各种特征，这些特征被应用于计算对比学习的损失函数，以确保正例和负例的损失都被最小化。每个步骤都经历了大量研究工作，众多学者提出了许多优秀的对比学习方法，如 InstDisc^[56]、MOCO^[57]、BYOL^[58]、CPC^[59]、CMC^[60]、SimCLR^[61]等。

2.5.1 常用的对比学习框架

本节主要讨论几种典型常用的对比学习方法。

(1)InstDisc 提出了个体判别的代理任务，引入了 memory bank 数据结构来储存大量负样本，作为 MoCo 的前身。将每张图片视为一个独立的类别，旨在让模型学习图像的特征，以实现对各种图像的区分。

(2)SimCLR 是一种基于负例的对比学习方法,其训练流程分为两个阶段:数据增强和对比学习。该框架在行人重识别任务中展现出显著优势,能够有效挖掘跨摄像头场景下的行人身份一致性特征。

数据增强阶段:通过序列化应用色彩空间变换(如亮度调整、饱和度扰动)、几何形变(随机裁剪、水平翻转)以及遮挡模拟(矩形区域擦除)等操作,生成具有视角差异性的增强样本。这种增强机制在行人重识别场景中可有效模拟跨摄像头拍摄差异,例如不同监控视角下的光照条件变化、行人姿态偏转及部分遮挡等情况。为进一步提升对复杂场景的适应性,改进方案可引入人体姿态引导增强,基于关键点检测结果针对性生成肢体运动模糊样本,强化模型对姿态变化的鲁棒性。

对比学习阶段:从每个样本的 K 个增强样本中随机选择两个,将其视为同一图像在不同视角下的表示。经过特征编码器和非线性投影层(Projection Head)的处理后,样本被映射到低维度的表示空间。模型采用参数共享的对称双流拓扑架构,两个分支分别处理这两个视角的图像,通过最小化表示向量之间的距离,使得同一图像在不同视角下的表示尽可能相似。

SimCLR 框架首次证明了复杂的图像增强技术的重要性,同时强调了 Projection Head 结构的关键性。这两个因素的结合极大地提升了对比学习系统的性能,使其在某些情况下达到甚至超过了有监督学习模型。

(3)MoCo 是一种基于负例对比学习方法,框架创新性地构建了动态队列维护的动量编码器体系,其核心突破在于采用非对称双分支架构实现大规模负样本的高效对比学习。在两个网络之间建立一种“查询器”和“记忆库”的关系。查询器负责生成特征向量,而记忆库则存储了图像特征的表示,并在每次迭代中逐步更新。这种方法允许查询器与记忆库中的特征进行对比,从而形成有效的负样本集合。通过这种方式,MoCo 不仅能够在未标记数据上有效学习到更具代表性的特征,同时也避免了对比学习中的负样本数量不足问题,从而提升了在下游任务中的表现。

MoCo 使用两个神经网络对数据进行编码:编码器(encoder)和动量编码器(momentum encoder)。编码器主要职责是对当前输入实例进行特征提取与转换,从而生成该实例的特征表示,而动量编码器功能则更为广泛,它能够对包含当前实例在内的多个实例进行特征编码,进而生成这些实例的综合特征表示。在 MoCo 中,编码特征 key 是通过一定策略得到的,并且无需梯度更新。更新过程仅依赖于编码器,以调整输出的 key。

(4)BYOL(Bootstrap Your Own Latent)是一种自监督对比学习框架,其重点在于在线更新目标网络参数。该框架采用一对在线网络和目标网络,通过在线网络权重的指数移动平均值来更新目标网络。BYOL 专注于学习表示,无需考虑反例。

这一框架致力于最大程度地增强同一实例的不同视图之间的一致性,同时将相似性估计与负面示例解耦。BYOL 在计算机视觉、自然语言处理等多个领域展现出令人瞩目的性能。它在表示质量方面取得显著进展,并在多个基准数据集上达到了最先进水平结果。

2.5.2 图像增强

在深度学习模型的训练阶段,过拟合是一个较为棘手的问题。一旦模型出现过拟合,就会过度“记住”训练样本的特征,从而在面对测试样本时,无法准确识别和处理,丧失了应有的泛化能力。而数据增强技术,能够在一定程度上有效缓解这一难题。

在对比学习领域,数据增强更是不可或缺的关键环节。这与普通深度学习任务有所不同,在普通任务里,数据增强并非必选操作,具有一定的可选择性。常见的图像处理手段,像裁剪、遮挡、旋转、添加噪声、调整对比度、调节亮度以及模糊处理等,都常被运用到数据增强的流程当中。经过这些处理后,图像会呈现出各异的变化效果,为模型训练提供了丰富多样的样本,助力模型学习到更广泛的特征,提升其泛化性能。



图 2-12 常见数据增强效果图

Fig. 2-12 Common data enhancement effect diagram

Chen 等^[61]系统地研究了不同数据增强方法以及它们之间组合方式对对比学习模型分类准确率的影响。研究表明:单独使用任何一种数据增强方式的性能提升有限;而将随机裁剪与色彩失真方法结合使用可以取得最佳效果。

第 3 章 结合随机遮挡和对比裁剪的无监督行人重识别

行人重识别旨在识别不同摄像机视角下的同一个人,是模式识别和计算机视觉中一项流行且重要的任务,它已被广泛应用于大规模监控系统中。该技术对天网工程、智慧城市的推进建设具有重大意义。由于深度卷积神经网络具有卓越的特征提取能力。考虑到行人重识别的识别精度常受到遮挡、光照、行人姿态等影响。因此,如何提取有效的行人图像信息,提升行人重识别的准确性,非常具有挑战性。

如今深度学习算法取代了传统的度量学习方法,在当今的行人重新识别中占据了主导地位。Wang 等^[62]提出一种多分支模型(MGN),能较好地对不同粒度下的具有表征能力的深度特征进行提取。Yan 等^[63]提出一种轻量级的基于局部的表示增强网络,利用局部相关性来聚集局部特征的独特信息。Wang 等^[64]提出一种基于变换器的姿态引导特征分离方法,利用姿态信息来清晰地分离语义成分,并选择性地匹配非遮挡部位。库等^[65]在残差网络的基础上,引入 CBAM 注意力机制模块^[66]用于处理遮挡情形下的行人重识别,增强有用的特征和抑制作用不大的特征来提升网络的判别能力。南等^[67]融合行人图像的多种信息,提出一种基于 U-Net 的软掩膜前景分割算法并且结合多信息融合重排序算法,缓解了背景变化对于行人重识别的负面影响。汪等^[68]利用 CNN 和 Transformer 特征学习网络的特点,提出了一种全局-局部融合特征的遮挡行人重识别方法。周等^[69]提出关键细粒度信息指导的多尺度遮挡行人重识别网络,减轻了背景和遮挡等干扰信息对行人重识别任务的干扰。

以上有监督的行人重识别方法已经取得了优秀的成果,但由于有监督学习领域过度依赖行人标注,且随着监控数据的海量增长使得计算资源成本也不断增加,从而限制了行人重识别技术在实际应用中的进一步扩展。基于以上原因,如何利用无需标记数据的无监督行人重识别方法,受到越来越多的关注。为了缓解行人重识别任务对标注数据依赖和遮挡场景下准确率低的问题,本文提出一种结合随机遮挡和对比裁剪的无监督行人重识别方法,从多个维度来优化 re-id 网络,对不同粒度信息进行更加深度的特征提取,能够使样本在特征空间中达到更好的聚类效果,缓解梯度消失的问题并且增强模型的泛化能力,有效提升遮挡场景下行人重识别的准确率。

3.1 方法具体内容

3.1.1 方法概述

本文方法涉及的主要内容有：随机遮挡^[70]、对比裁剪(ContrastiveCrop)^[71]、实例间对比编码(ICE, Inter-instance Contrastive Encoding)^[72]以及 SC-CAResNet 提取网络几个部分。整体方法框架图如图 3-1。

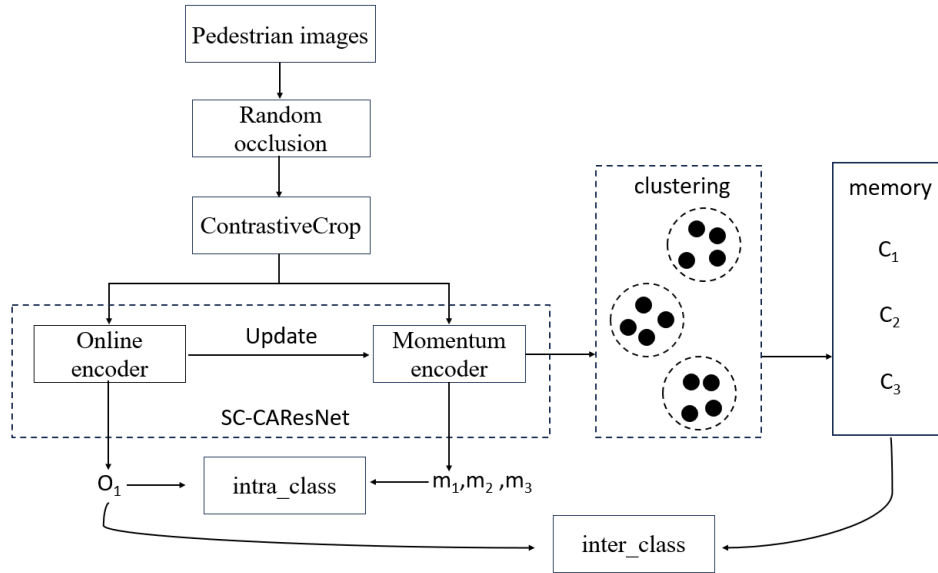


图 3-1 方法框架(不同簇的中心点 C_1, C_2, C_3 ; 经过编码器得到特征向量 O_1 和 m_1, m_2, m_3)

Fig. 3-1 Method framework. (The center points of different clusters C_1, C_2, C_3 ; after the encoder operation, the feature vectors O_1 and m_1, m_2, m_3 are obtained.)

首先引入随机擦除模块对原始输入图像进行随机遮挡生成行人受遮挡图像，对遮挡图像使用对比裁剪，对比裁剪不同于常规的随机裁剪(RandomCrop)，对比裁剪通过对比学习方法，由中心压制的非均匀分布进行采样，增加多次采样的方差，加大了不同样本之间的差异，将这两种方法与基准模型 ICE 联合使用。该方法是 MoCo 模型优化版，在 MoCo 模型的基础上添加了聚类算法，可以为样本生成类别标签，以此计算类间损失和类内损失，得到更加紧凑的聚类结果。其中的动量编码器(Momentum encoder)和在线编码器(Online encoder)是基于 ResNet50 网络设计的，两者结构相同，动量编码器(权重设为 θ_m)，由在线编码器(权重设为 θ_o)更新：

$$\theta_m^t = \alpha \theta_m^{t-1} + (1 - \alpha) \theta_o^t \quad (3-1)$$

式中 α 是动量系数， t 表示当前迭代。

在每一轮训练开始时使用动量编码器提取训练集得到特征向量，对其使用聚类算法 DBSCAN^[73]生成伪标签。通过公式(3-2)计算质心 C_1, C_2, C_3 。

$$p_a = \frac{1}{N_a} \sum_{c_i \in \mathcal{Y}_a} c_i \quad (3-2)$$

利用在线编码器、动量编码器计算出的特征向量,并结合聚类产生的类信息,计算类间损失和类内损失^[72]。

考虑到编码器中 ResNet50 网络难以提取多粒度的行人特征,将其中的常规卷积替换为空间和通道重建卷积(Spatial and Channel Reconstruction Convolution, SCConv)^[74]加强特征的适应性,并在残差结构中添加坐标注意力机制(Coordinate Attention, CA)^[75]优化特征提取,设计出 SC-CAResNet 网络,更能针对性地提取关键特征,关注行人重识别任务中需要的多粒度信息,两者结合引导网络同时考多个维度,使网络更加全面关注遮挡场景中行人的显著区域,缓解梯度消失的问题并且增强泛化能力,最后联合使用类间损失、跨镜头差异损失、难样本对比损失和一致性损失对网络进行训练,改善样本采集时镜头差异,数据分布变化等风险,缓解遮挡场景下行人重识别精度不高的问题。

3.1.2 随机遮挡

近年来,深度学习为计算机视觉领域的进步做出了重大贡献,大多可归因于卷积神经网络的使用,它能够学习图像的复杂层次特征表示;但随着待解决任务的复杂性增加,此类模型的资源利用率也会增加。目前网络通常包含数量庞大的学习参数,为了提高视觉任务表示能力,这类参数必不可少。但随着表示能力的增加,过拟合的可能性也会增加,导致泛化能力差。

随机擦除是一种易于实现的数据增强方法,属于信息删除方法的一种,不仅降低了过拟合的风险概率,而且提升了网络模型应对遮挡问题的泛化能力^[70],经过随机擦除这一数据处理方法模拟现实场景中图像部分信息被遮挡的情况,图像在某个区域的特征无法被识别获取。对于此类方法,首先要解决的是如何避免对输入图像信息过多或过少的删除。遮挡区域过大或过小都不利于网络提取性能,会对网络的鲁棒性和泛化能力产生负面影响。大多遮挡操作是在图像上随机框出一个规则区域进行随机擦除。这样虽然能取得较好的效果,但在实际场景中,遮挡物多样且形状不规则。

本文设置随机遮挡行人图像的概率为 50%,在整个输入图像上随机框出一个矩形和一个圆形区域进行随机遮挡,当两个遮挡图形区域重合时可以组合成多种随机图案。更加符合现实场景的遮挡情形,随机遮挡区域相对于输入图像的比例范围为 0.02~0.4。这可以更好地为模型准备好应对现实世界中的遮挡情况。对行人图像的遮挡效果如图 3-2 所示。



(a) 行人原始图像



(b) 使用规则图形遮挡行人图像



(c) 使用不规则图形遮挡行人图像

图 3-2 行人图像

Fig. 3-2 Pedestrian images

3.1.3 对比学习

对比学习是一种自监督学习方法，常用于在没有标签的情况下，通过让模型学习数据点相似或不同来学习数据集的一般特征，在许多的下游任务中被认为是有效的。随机裁剪就是对比学习种常用的图像增强操作之一，如式(3-3)所示：

$$(x, y, h, w) = \mathbb{R}_{crop}(s, r, I) \quad (3-3)$$

式中， s 表示剪裁大小， r 表示剪裁高宽比， I 表示输入的行人图像， R_{crop} 表示随机裁剪函数，得到一个四元组 (x, y, h, w) ，代表了剪裁的中心坐标，高度和宽度。但这种直接的裁剪可能会生成错误的正样本对，例如图 3-3(a)。或者生成过于相似的正样本对，这些正样本难以提升模型识别精度，如图 3-3(b)。尤其在行人受遮挡情境下，常常会把遮挡物裁剪出来当作正样本。因此设计一种遮挡场景下可以通过辨别行人图像语义信息的裁剪方法进行图像增强是必要的。

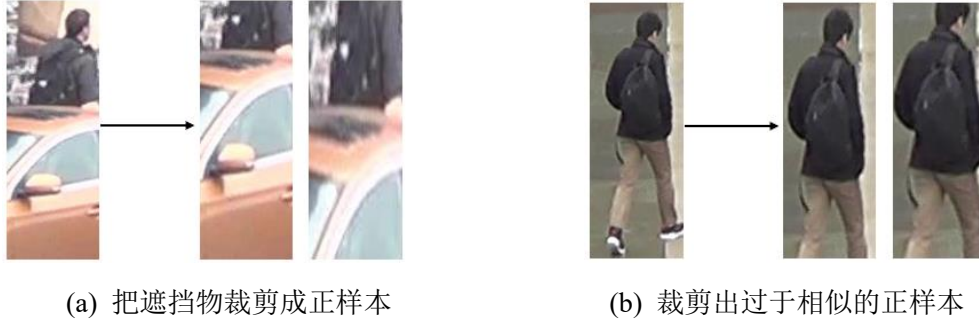


图 3-3 随机裁剪的行人图像

Fig. 3-3 Randomly cropped pedestrian images

考虑到数据增强操作在对比学习中的重要性，为了避免上述的裁剪错误，Peng 等^[71]进行了更加深入的研究，使用一种语义感知的方法，通过加入定位信息来减少随机裁剪过程中产生的错误样本对。首先根据训练过程中的特征图来获取行人大致区域，如公式(3-4)所示：

$$B = L([M > k]), k \in [0, 1] \quad (3-4)$$

其中 M ， k ， 1 ， L 分别代表热度图、激活点的阈值、指示函数和计算矩形闭包的函数，以上内容返回一个定位框 B 。在得到定位框 B 后，采样可以被表示为：

$$(\dot{x}, \dot{y}, \dot{h}, \dot{w}) = \mathbb{R}_{crop}(s, r, B) \quad (3-5)$$

此外，为了裁剪出具有更大差异的正样本对，获得更全面的全局视野和细粒度的局部信息，引入“中心压制采样”的方法，通过两个相同参数 α 的 Beta 分布 $\beta(\alpha, \alpha)$ 作为采样的概率分布，通过 $\alpha < 1$ 的操作，使分布呈现中间低、周围高的 U 型。这种方法不仅能够降低裁剪区域集中在图像中心的概率，还能够减少裁剪区域间的相互重叠。对比裁剪最终由公式表示为：

$$(\dot{x}, \dot{y}, \dot{h}, \dot{w}) = \mathbb{C}_{crop}(s, r, B) \quad (3-6)$$

其中 $\mathbb{C}_{crop}$ 表示中心压制采样的函数。图 3-4 展示了随机裁剪和对比裁剪的差异对比。

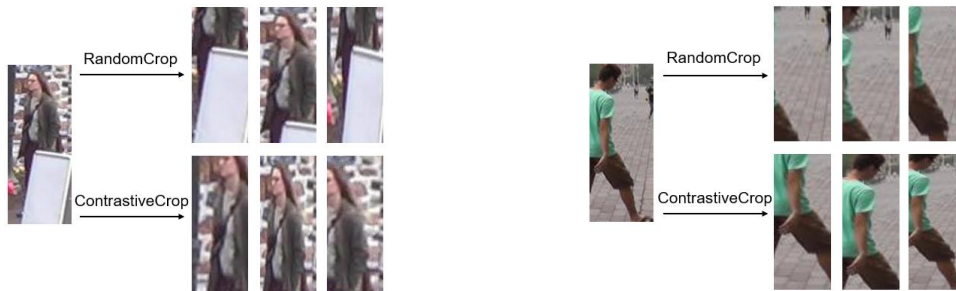


图 3-4 不同裁剪方式对比

Fig. 3-4 Comparison of different cutting methods

根据上图所示,引入的对比裁剪的方法,通过识别行人图像中关键的语义信息来进行裁剪图像的生成,不仅确保了大部分正样本对语义一致,而且加大了行人样本之间的差异性,使网络学习到更泛化的行人特征。可以有效应对遮挡场景和杂乱背景的干扰,在稳定提升模型识别精度的前提下而不造成更多额外的训练和计算代价。

3.1.4 SC-CAResNet 网络模型

近年来随着 Vision Transformer(ViT)模型的应用,凭借其优越的自注意力机制网络设计,为计算机视觉领域提供了崭新的思路和方法,很多研究者将其视为卷积神经网络的替代品。但是 ViT 模型有更大的计算开销,相较于 ViT 模型,卷积神经网络运算训练成本更低,能够有较好的硬件支持,仍然在很多轻量级模型领域占据主要地位,在行人重识别领域发挥着重要的作用。受文献^{[74][75]}的启发,在保证更广泛的硬件支持和更高的推理速度的前提下,本文设计出新的提取网络,该网络使用空间和通道重建卷替换 ResNet50 中标准卷积并且加入坐标注意力机制,新网络称为 SC-CAResNet。该网络能在保持训练速度的前提下,充分关注行人图像的局部信息和全局信息,拥有较高的识别精度,更好的应用于行人重识别任务。

空间和通道重建卷积模块插入位置如图 3-5 所示,由两部分组成:空间重建单元(spatial reconstruction unit, SRU)和通道重建单元(channel reconstruction unit, CRU)以顺序串联的方式执行。

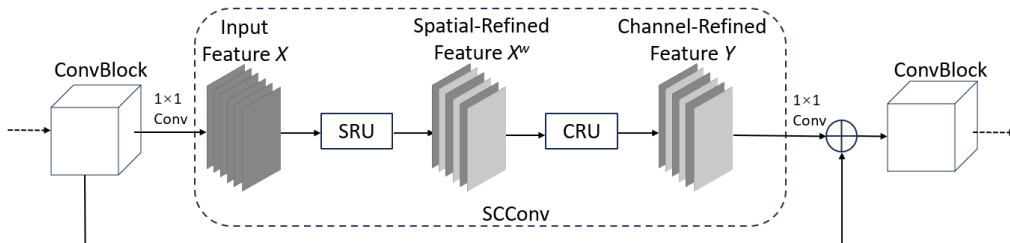


图 3-5 SCConv 插入位置

Fig. 3-5 SCConv insertion position

输入特征 X 首先经过空间重建单元处理得到具有空间细化特征 X'' , 然后再通过通道重建单元处理, 获得通道细化特征 Y 。两个模块联合使用可以有效增强 CNN 网络的特征表示并减少特征之间空间和通道的冗余。空间重建单元用于处理特征的空间冗余, 空间重建单元用于处理通道冗余。其具体处理过程如图 3-6 所示。

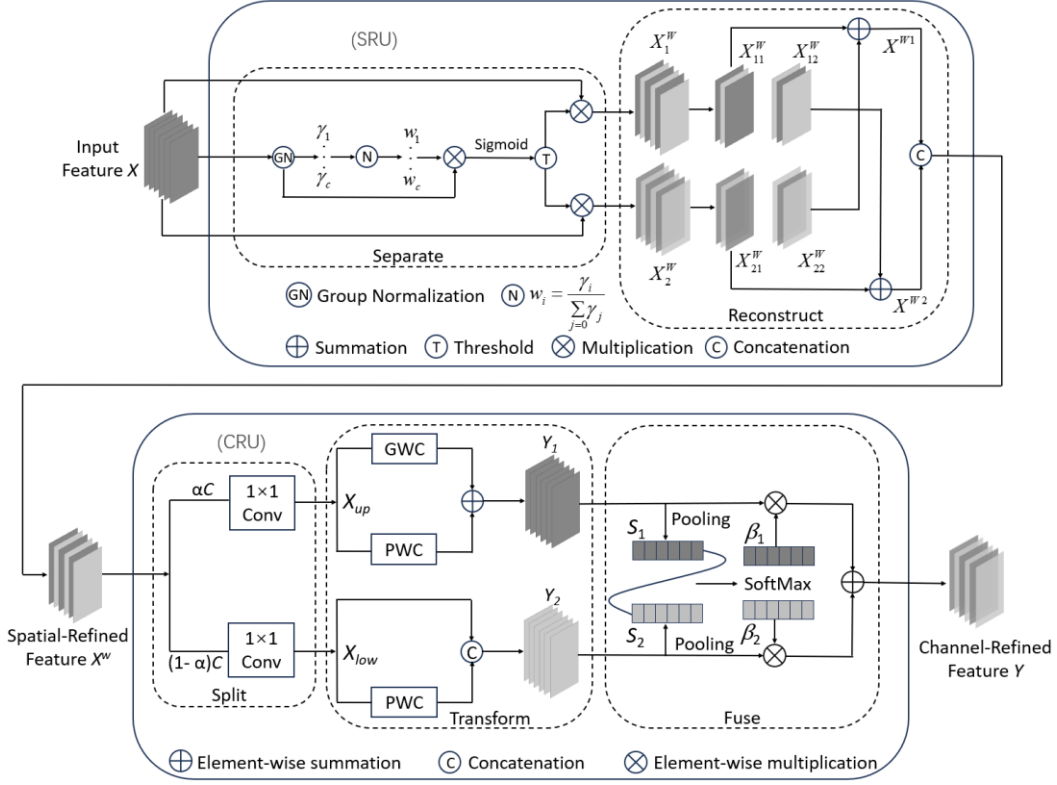


图 3-6 SCConv 处理流程

Fig. 3-6 SCConv processing flow

空间重建单元分为分离和重构。分离操作首先使用组归一化层来评估不同特征图的信息内容，对中间特征映射 $X \in \mathbb{R}^{N \times C \times H \times W}$ 减去平均值 μ 同时除以标准差 σ 得到标准化的输入特征 X ，此外，一个大于 0 的小常数 ε 和两个可训练的仿射变换参数 γ ， β 均有助于公式算法的稳定性，如公式(3-7)所示：

$$X_{out} = GN(X) = \gamma \frac{X - \mu}{\sqrt{\sigma^2 + \varepsilon}} + \beta \quad (3-7)$$

之后利用 Group Normalization 操作训练得到 γ 用于测量每个批次和通道的空间像素方差， w_γ 为归一化相关权重反映不同特征图的重要程度，其计算方法如公式(3-8)：

$$W_\gamma = \{w_i\} = \frac{\gamma_i}{\sum_{j=1}^C \gamma_j}, i, j = 1, 2, \dots, C \quad (3-8)$$

将经过处理的权值使用 sigmoid 函数映射到(0,1)范围，通过阈值进行门控，得到信息权重 w_1 ，非信息权重 w_2 。计算方法如公式(3-9)：

$$W = Gate(Sigmoid(W_\gamma(GN(X)))) \quad (3-9)$$

最后将 W_1 和 W_2 分别和特征 X 相乘,得到含有丰富信息的 X_1^w 和信息较少的 X_2^w 。重构操作则是将具有不同信息量的特征交叉重构,得到更丰富信息的特征 X^w ,过程如公式(3-10):

$$\begin{aligned} X_1^w &= W_1 \otimes X, \\ X_2^w &= W_2 \otimes X, \\ X_{11}^w \oplus X_{22}^w &= X^{w1}, \\ X_{21}^w \oplus X_{12}^w &= X^{w2}, \\ X^{w1} \cup X^{w2} &= X^w \end{aligned} \quad (3-10)$$

此外为了利用特征的通道冗余,使用分割变换和融合策略来构建通道重建单元。通过分割、转换和融合三部分操作实现。第一步是分割,将上一层空间重建单元的输出特征 X^w 进行分割得到通道数分别为 αC 和 $(1-\alpha)C$ 两个部分($0 \leq \alpha \leq 1$),随后使用 1×1 卷积来压缩特征映射的通道,得到 X_{up} 和 X_{low} 。第二步是转换,将 X_{up} 作为“丰富特征提取器”,使用高效的卷积运算: Group-Wise Convolution(GWC)和 Point-Wise Convolution(PWC),结果相加得到代表性特征图 Y_1 ,如公式(3-11):

$$Y_1 = M^G X_{up} + M^P X_{up} \quad (3-11)$$

X_{low} 进行 PWC,和原本的输入取并集得到 Y_2 。如公式(3-12):

$$Y_2 = M^P X_{low} \cup X_{low} \quad (3-12)$$

第三步是融合,利用简化的 SKNet 合并上一层输出的 Y_1 和 Y_2 。得到经过池化操作的 S_1 和 S_2 ,计算方法如公式(3-13):

$$\begin{aligned} S_m = Pooling(Y_m) &= \frac{1}{H \times W} \sum_{i=1}^H \sum_{j=1}^W Y_m(i, j) \\ m &= 1, 2 \end{aligned} \quad (3-13)$$

之后由 S_1 和 S_2 经过通道软注意力操作得到特征权重向量 β_1 和 β_2 。最后使用公式(3-14)经行合并输出处理的最终特征 Y :

$$Y = \beta_1 Y_1 + \beta_2 Y_2 \quad (3-14)$$

注意力机制最初应用于机器翻译任务,现在已被广泛地应用于深度学习的各个领域例如:图像分类、目标检测和分类等。常用的注意力机制包括 SE 注意力机制^[76]和 CBAM 注意力机制^[66]等。

本文使用的坐标注意力机制能够同时考虑通道间关系以及长距离的位置信息。它不仅获取了通道间信息，高效地处理通道间的关系，还包含了方向相关的位置信息，完全利用提取的位置信息，这有助于模型更好地定位和识别目标。坐标注意力机制结构图如下。最主要的操作是将全局池化变为两个 1D 的特征编码。

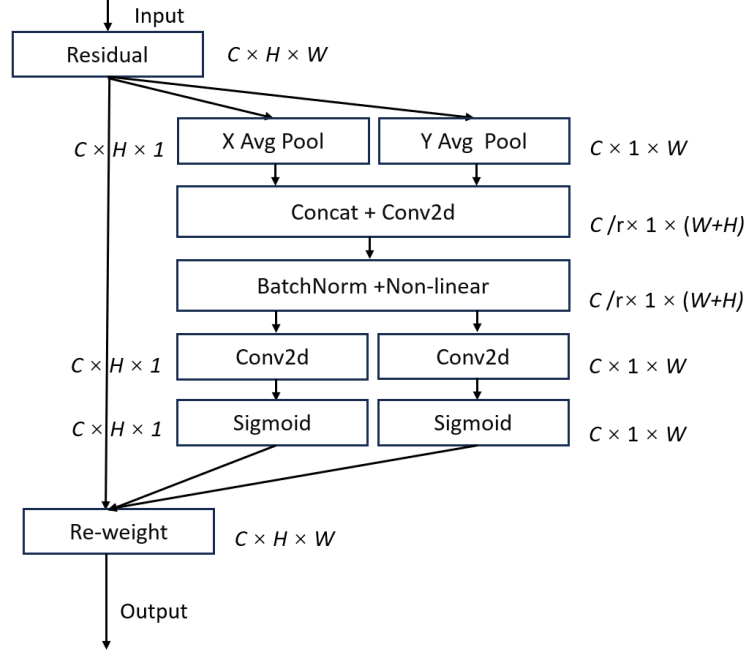


图 3-7 坐标注意力流程图

Fig. 3-7 Coordinate attention flowchart

本文将该注意力嵌入到 ResNet50 的 conv2、conv3 和 conv4 后部，提升网络的精度。这种注意力机制可以给网络处理任务带来明显性能提升。这种方法通过区分空间方向(即坐标)，并生成具有坐标感知的特征图，因此被研究人员命名为坐标注意力。相较于之前的注意力操作，坐标注意力能够充分利用提取的位置信息，加强通道之间的联系。

3.1.5 损失函数

类间对比损失定义如下，目的是让同类样本相似度变得更高，不同类样本相似度变得更低。

$$L_{inter_class} = \mathbb{E} \left[-\log \frac{\exp(o_a \cdot c_a / \tau_{inter_class})}{\sum_{i=1}^{|p|} \exp(o_a \cdot c_i / \tau_{inter_class})} \right] \quad (3-15)$$

其中， O_a 表示属于类 a 的特征向量， C_a 表示类 a 的中心点， τ_{inter_class} 是其中的一个超参数， $|p|$ 表示训练时的样本类别数。

跨镜头差异损失定义如下，目的是拉近同一类中不同镜头样本之间的距离，减少类内相机风格差异。

$$L_{camera_loss} = \mathbb{E} \left[-\frac{1}{|p|} \sum_{i \neq b \cap i \in C} \log \frac{\exp(s(o_{ab}, c_{ai}) / \tau_{camera})}{\sum_{j=1}^{Nneg+1} \exp(s(o_{ab}, c_j) / \tau_{camera})} \right] \quad (3-16)$$

其中 $|p|$ 表示跨镜头的正样本中心点数量，相对应的 $Nneg$ 则表示与当前样本距离最近的负样本， $s(\cdot)$ 表示余弦相似度， O_a 表示属于类 a ，但由镜头拍 b 摄的当前样本。

难样本对比损失定义如下，目的是减少聚类过程中离群点的产生，加强簇的紧凑性。

$$L_{hard} = \mathbb{E} \left[-\log \frac{\exp(s(o^i, m_k^i) / \tau_{hard})}{\sum_{j=1}^{J+1} \exp(s(o^i, m_j) / \tau_{hard})} \right] \quad (3-17)$$

其中， O^i 表示给定第 i 类的锚点， m_k^i 表示与 O^i 最相似的正样本， j 表示对于当前锚点不属于第 i 类的负样本数量。

为了防止数据增强后实例间的相似性可能会发生变化，保证在进行数据增强前后数据分布的一致性。引入一致性损失。

数据增强后的分布：

$$P_1 = \frac{\exp(s(O_A, m) / \tau_{aug})}{\sum_{j=1}^N \exp(s(O_A, m_j) / \tau_{aug})} \quad (3-18)$$

通过 KL 散度拉近数据增强前后分布的距离：

$$L_{aug} = \mathcal{D}_{KL}(P_1 \| P_2) \quad (3-19)$$

将以上损失函数联合使用，总体损失函数如下：

$$L = \lambda_1 L_{inter_class} + 0.5 L_{camera_loss} + \lambda_2 L_{hard} + \lambda_3 L_{aug} \quad (3-20)$$

为验证本文方法的有效性，本文在公开的两个行人重识别数据集 Market-1501 和 DukeMTMC-reID 上进行了验证。

实验采用 Rank-k 和平均精度(mAP)来验证本文所提方法。Rank-k 表示按照相似度排序后的前 k 张图像中，存在与查询图像属于同一 ID 的概率。本文实验中选取 Rank-1 作为评价指标，即第 1 张图片与查询图像属于同一个。mAP 表示平均精度均值，它能更加全面地衡量行人重识别算法的性能，衡量的是学出的模型在所有类别上的好坏，是将平均精度(AP)求和再取平均。

3.1.6 消融实验

为了验证本文所提出方法，在两个通用数据集 Market-1501 上进行消融实验，实验结果如表 3-1 所示。

表 3-1 在数据集上的消融实验

Table 3-1 Ablation experiments on the dataset

方法	Market-1501	
	Rank-1(%)	mAP(%)
Baseline	90.4	78.6
+遮挡	91.8	81.7
+对比裁剪	91.3	80.6
+SC-CAResNet	94.6	84.4
本文	95.6	85.7

由表 3-1 可以看出, 使用 Baseline 时的 Rank-1 和 mAp 为 90.4%, 78.6%。在此基础上分别加上遮挡, 对比裁剪, SC-CAResNet 这三个模块, Rank-1 指标分别达到了 91.8%, 91.3%, 94.6%; mAp 指标分别达到了 81.7%, 80.6%, 84.4%。由此看出单独添加随机遮挡、对比裁剪和 SC-CAResNet 网络均能提高行人重识别的性能。三种方法联合使用效果最好, 最终的 Rank-1 为 95.6%, mAp 为 85.7%。证明了本文提出的方法在遮挡场景下可以有效提高行人重识别的识别精度。

3.1.7 对比实验

为了进一步证明本文提出的方法能有效提高遮挡场景下行人重识别的精度, 改善模型的鲁棒性和泛化性。在两个通用数据集 Market-1501、DukeMTMC-reID 上, 将本文的方法与当今主流的方法进行性能对比, 各方法 Rank-1 及 mAP 如表 3-2 所示。

表 3-2 不同先进方法在数据集上结果对比

Table 3-2 Comparison of results of different advanced methods on the data set

方法	Market-1501		DukeMTMC-reID	
	Rank-1(%)	mAP(%)	Rank-1(%)	mAP(%)
HG ^[77]	95.6	86.1	87.1	77.5
ICE ^[72]	93.8	82.3	83.3	69.9
MCC-DCL ^[78]	94.8	85.8	84.8	74.4
PAN ^[79]	94.9	87.6	89.5	78.2
HQP ^[80]	82.3	80.3	82.6	68.0
OCNet ^[81]	95.0	89.3	90.5	80.2
MHSA-Net ^[82]	94.3	82.5	87.0	72.6
本文	95.6	85.7	86.7	76.0

从表 3-2 中可以看出, 在 Market-1501 数据集上 Rank-1 和 mAP 指标分别达到了 95.6%和 85.7%。在 DukeMTMC-reID 数据集上 Rank-1 和 mAP 指标分别达

到了 86.7%和 76.0%。在两个数据集中，本文方法的指标多数处于前列。对比实验表明，本文提出的方法能够有效提升行人被遮挡时的识别能力。

此外，为了进一步研究使用规则图形遮挡行人图像和不规则图形遮挡行人图像得到的最后识别效果，在数据集 Market-1501 上进行对比，结果如表 3-3 所示。

表 3-3 不同遮挡方法的对比

Table 3-3 Comparison of different occlusion methods

方法	Market-1501	
	Rank-1(%)	mAP(%)
规则图形遮挡	91.5	81.2
不规则图形遮挡	91.8	81.7

结果显示，使用不规则图形对图像进行遮挡最后实验效果优于使用规则图形对图形进行遮挡。

3.2 本章小结

无监督行人重识别相对于有监督行人重识别具有不依赖行人真实标签数据，节约数据标注成本的优点，本文对遮挡场景下的无监督行人重识别进行研究，提出一种结合随机遮挡和对比裁剪的无监督行人重识别方法。

首先对数据集进行概率为 50%的随机遮挡，在图像上随机框出多种随机图案生成行人受遮挡图像。其次引入的对比裁剪方法则可以保证正样本对语义一致的前提下，尽量裁剪出行人主体部分，生成差异更大的行人图像样本，缓解遮挡和复杂背景对网络的影响；设计出 SC-CAResNet 替换对比学习方法 ICE 中常规 ResNet50 强化了网络的特征提取能力；最后利用多种损失函数进行联合损失函数训练。通过消融实验和对比分析，结果表明本文方法在提升网络识别精度的前提下，保证了模型的鲁棒性和泛化性。

实验在两个主流数据集上进行训练和评价，取得了良好的性能，达到了有监督时的准确率，反映了所提方法具有较强的竞争力。

第 4 章 基于暗光增强和细粒度特征提取的行人重识别

在 2017 年, Vaswani 等提出的 Transformer 模型首先应用于自然语言处理, 因为其优异的性能和简单的模型结构迅速引起众多学者关注; 随后 Dosovitskiy 等^[8]提出了 ViT(Vision Transformer, ViT)模型, 将 Transformer 引入了 CV 图像领域, 为计算机视觉领域带来新的思路和方法; He 等^[46]首次将 Transformer 用于行人重识别和车辆重识别并提出拼图补丁模块(Jigsaw Patch Module, JPM)和侧信息嵌入(Side Information Embeddings, SIE)取得了较好的性能; Luo 等^[83]提出 BNNeck, 可将交叉熵损失和三元组损失更好地联合使用于行人重识别。

Transformer 方法在视觉任务中展现了巨大的潜力, 逐步显示出替代卷积神经网络的趋势。然而, 纯 Transformer 网络缺乏归纳偏差, 计算复杂度较高, 训练过程较慢, 同时对硬件的需求也较大。

针对以上问题, 考虑到现实场景下不同光照会使行人图像出现颜色, 纹理等误差, 本文提出一种基于暗光增强和细粒度特征提取的行人重识别方法。首先使用暗光增强模块 IceNet^[84]对部分行人图像中的暗光区域增亮, 这种图像增强方法不仅能在多种场景下表现优异, 而且可以根据使用者需求, 自适应生成增强后的行人图像; 之后将新型的位置编码: 条件位置编码 CPE^[85]和拼图补丁 JPM 同时嵌入 Transformer 中, 提取更加鲁棒的行人特征; 最后受文献^[86, 87]启发, 设计出 DAD-T 模块对模型进行轻量化改造, 缓解模型性能下降, 过拟合的风险, 减少复杂度, 增强网络的表达能力。

4.1 方法具体内容

4.1.1 方法概述

基于 CNN 的方法难以利用全局范围内的行人信息, 往往局限于小的局部区域^[88]。注意力机制关注重要特征同时忽视次要特征, 虽然将注意力机制引入 CNN 网络, 给行人部件分配权重, 可以缓解远程依赖关系^[89], 周等^[90]提出了基于卷积注意力机制和多损失联合的跨模态行人重识别方法, 在 ResNet50 网络嵌入卷积注意力机制模块以抑制颜色等无关信息的提取。曾等^[91]设计混合池通道注意模块(HPCAM)和全像素空间注意力模块(FPSAM), 并基于这两种注意力模块提出一种通道与空间双重注意力网络(CSDA-Net), 但是注意力机制大多嵌入在深层, 没有从根本上解决原理问题。尤其是, 行人重识别属于复杂场景下的细粒度任务, 只靠单纯的 CNN 方法已经很难有大的突破, 需要设计特定的损失函数和修改网

络结构。最近，一些用 ViT 替代 CNN 的模型开始流行，ViT 在图像处理领域取得了优异的性能。

一般情景下，ViT 模型用于行人重识别时的网络框架图如图 4-1 所示，输入行人图像为 $x_i \in \mathbb{R}^{H \times W \times C}$ ， i 表示一个批量输入的第 i 张行人图像， H 、 W 和 C 分别表示图像的高度，宽度和通道数，首先将其切分为 N 个大小相等的图像块 $\{x_p^i | i=1, 2, \dots, N\}$ ，将图像块输入 Embedding 层，线性映射成 D 维向量，即每个图像块对应得到一个 D 维向量。值得一提的是，这一步操作主要通过卷积实现。

此外，ViT 在向量之前添加一个随机初始化可学习的分类标记([cls])。输出的[cls]标记用作全局特征表示 f ，在 f 之后使用 BNNeck 联合交叉熵损失和三元组损失。

输入到 Transformer 层的输入序列可以表示为：

$$Z_0 = [x_{\text{cls}}; F(x_p^1); F(x_p^2); \dots; F(x_p^N)] + P \quad (4-1)$$

式中， Z_0 表示输入序列嵌入， $P \in \mathbb{R}^{(N+1) \times D}$ 是位置嵌入， F 是将切割后的图像映射到 D 维的线性投影。

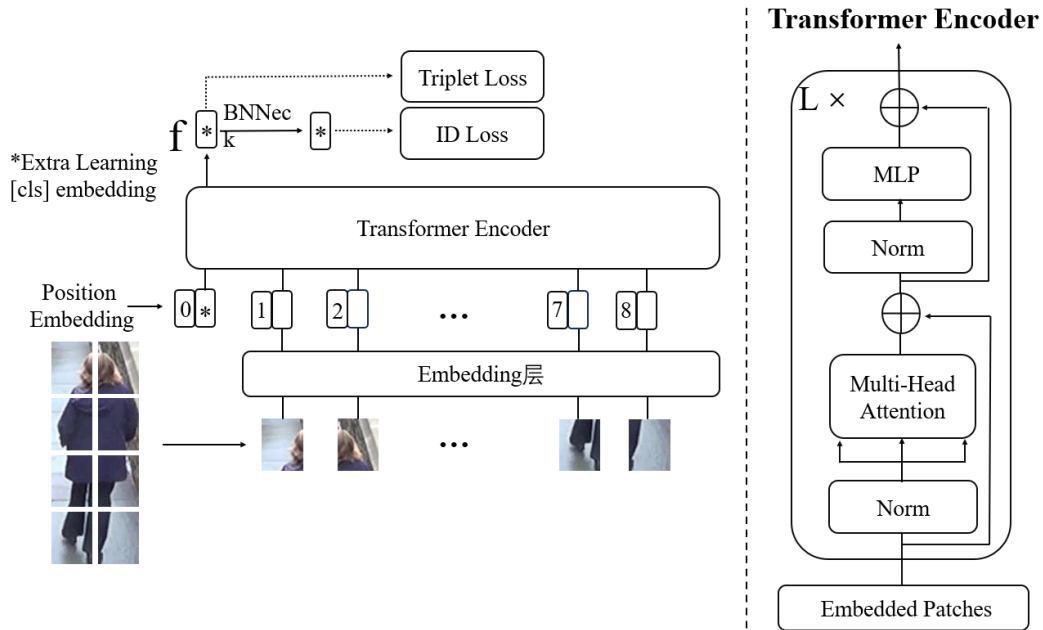


图 4-1 基于 ViT 的行人重识别网络框架图

Fig. 4-1 ViT-based person re-identification network framework diagram

4.1.2 改进的 ViT 行人重识别

以上基于 ViT 行人重识别模型虽然取得较好的成果，但如何让网络更关注包含行人特征的图像块，抑制现实场景下无用的背景信息，缓解因光照、遮挡等

导致的多种负面问题,本文设计了基于暗光增强和改进 ViT 的行人重识别模型,如图 4-2 所示,进一步提高行人重识别的准确性和识别鲁棒性。

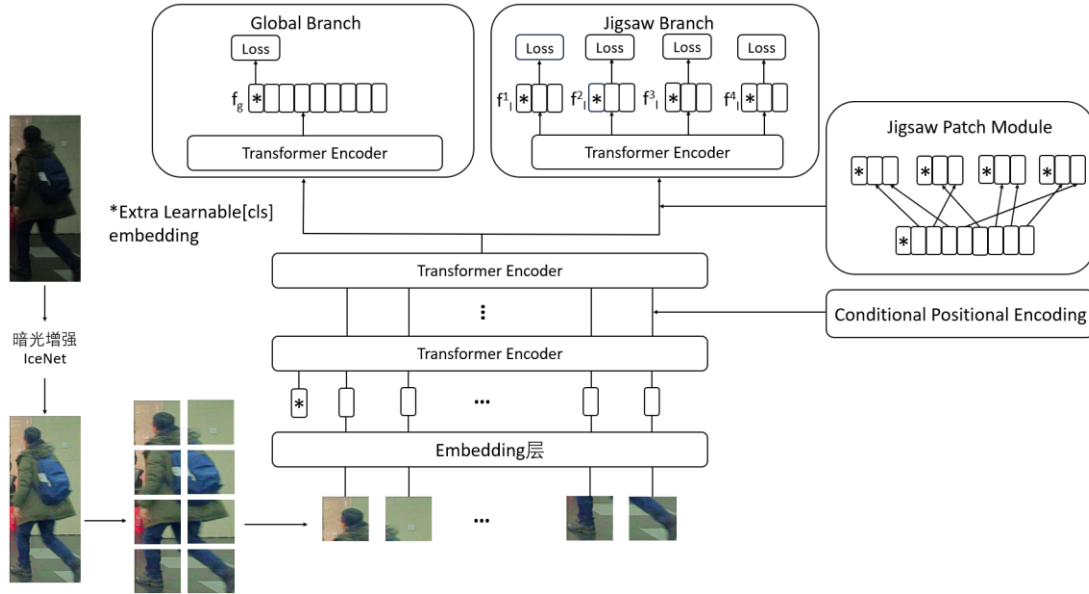


图 4-2 网络架构

Fig.4-2 Network architecture

输入图像首先经过 IceNet 网络进行图像增强,使行人图像更加清晰,色彩更加明亮,之后用滑动窗口,对其进行分块处理,在第一个 encoder 之后插入动态可生成的条件位置编码(CPE),由此泛化到更长的输入序列,可以和 CNN 的方法一样,拥有平移不变性,增加识别精度。同时由于行人重识别属于复杂场景下的细粒度任务,行人切片之间联系密切,每个局部片段如果仅考虑连续补丁嵌入的一部分,会造成资源浪费,由此使用拼图补丁模块(JPM),引入额外的扰动,也有助于提高目标 ReID 模型的鲁棒性。此外将 ViT 的最后一层改为 2 个独立的分支 Global Branch 和 Jigsaw Branch,以此来学习全局特征和局部特征。

4.1.3 暗光增强模块

光照、色度变化导致行人图像外观差异很大,难以完整的表达行人外观特征,这类图像对比度低,能见度差,严重干扰基于视觉的下游任务。暗光增强作为一个图像处理问题已经被研究了很长的时间。部分网络模型为了降低不同光照对模型的影响,使用灰色图像代替原始图像,但这类方法损失了图像的色彩信息。一些传统的低光照图像方法例如直方图均衡化通过灰度变换将一幅图像转换为另一幅均衡直方图,这类方法处理后的图像不仅灰度级减少,局部细节缺失,并且图像对比度变得异常增高。

本文在不改变原始图像结构信息的前提下,为了解决这些问题,引入暗光增强模块 IceNet 对输入图像进行处理,能够有效应对光照变化对模型的影响,提升模型适应视域变化的能力。该模块是一种基于 CNN 的交互式图像增强算法,可

以根据使用者的要求自适应地生成增强图像，也可以自动生成无需交互的图像。IceNet 非常灵活，框架图如图 4-3，适用于增强来自不同领域的真实低光图像。有益于加速网格模型的收敛，防止梯度消失，明显优于传统算法。

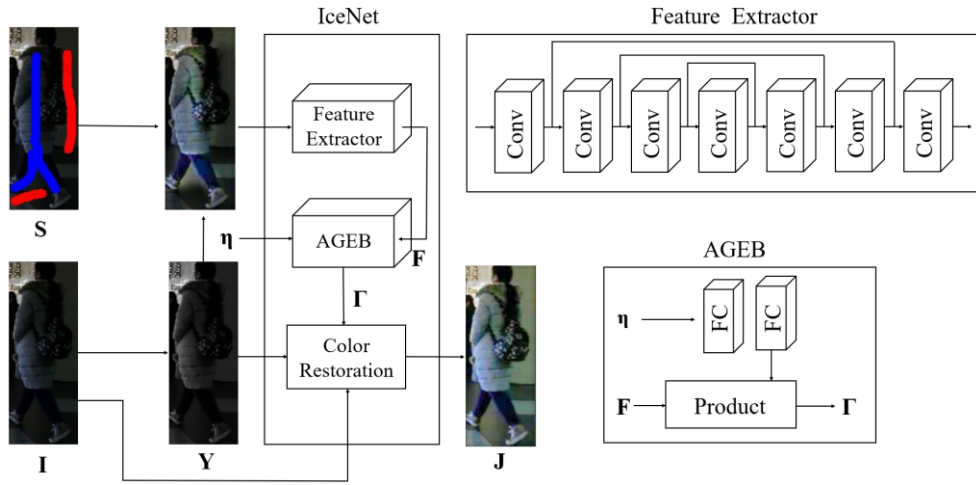


图 4-3 暗光增强流程图

Fig. 4-3 Dark light enhancement flow chart

增强效果如图 4-4，经过处理的行人图像更加清晰，色彩更加明亮。



图 4-4 增强效果对比图

Fig. 4-4 Enhancement effect comparison chart

4.1.4 条件位置编码

Transformer 中位置编码已经被证明是有效的，但会降低网络的灵活性，在 Transformer 中，没有对位置编码做特殊的设计，只采用一组可学习的参数来学习位置编码，该方法难以处理高分辨率图像。

Transformer 编码位置表示的方法主要有两种形式：绝对位置编码和相对位置编码。使用绝对位置编码的 ViT 灵活性和泛化性不佳，它将输入 token 的绝对位置从 1 编码到最大序列长度。也就是说，每个位置都有一个单独的编码向量。这些位置编码与输入的 token 一起结合，帮助模型理解每个 token 在序列中的相对位置，绝对位置编码会使得网络模型难以解决图像处理中的平移不变性。相对

位置编码通过编码输入 token 之间的相对距离，来学习 token 之间的相对关系。相对位置编码则会增加时间复杂度，加大模型变得训练难度。

本文采取一种条件位置编码，让 ViT 的输入更灵活。受 CNN 卷积的启发，在 ViT 的基础上，融合 CNN 平移不变性的优点，能更加适合处理较长地输入序列，提高行人重识别分类精度。该方法由一个简单的位置编码生成器(Position Encoding Generator, PEG)实现，它可以便捷地插入 ViT 模型中，本文在第一个 encoder 之后插入 PEG，两类位置编码的插入位置对比如图 4-5 所示。

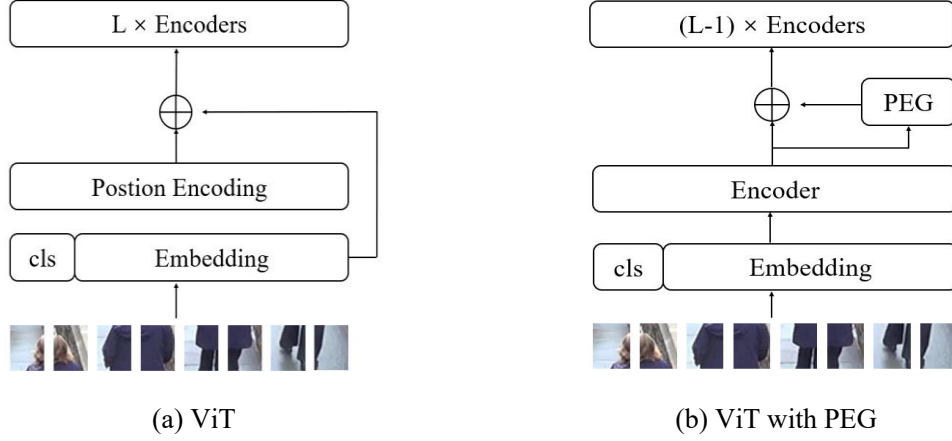


图 4-5 位置编码的插入位置对比图

Fig. 4-5 Comparison of insertion positions of positional coding

PEG 处理过程如图 4-6 所示。

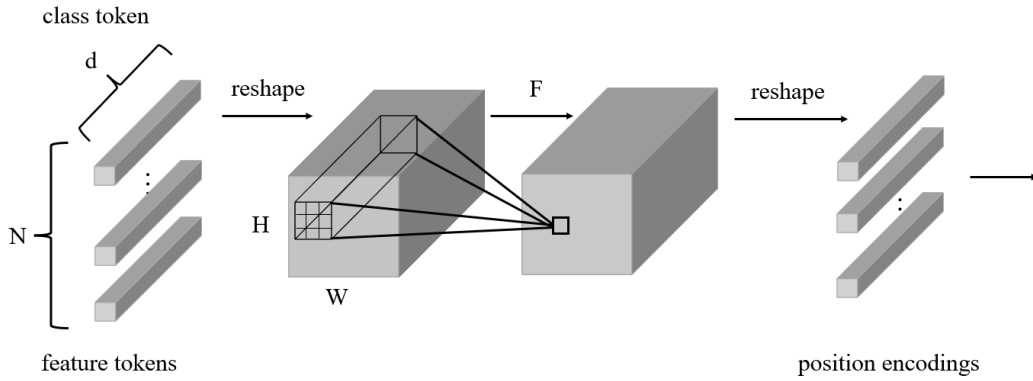


图 4-6 PEG 示意图

Fig. 4-6 PEG diagram

图中 PEG 的处理过程是将模型输入序列 $X \in \mathbb{R}^{B \times N \times C}$ 重塑为二维图像形状 $X' \in \mathbb{R}^{B \times H \times W \times C}$ ，其中 B 表示每批处理的图像数量； N 表示 token 数也就是将行人图像切分成的块数； d 表示每个块数的尺寸为 $d \times d$ ； C 表示图像的通道数； H 和 W 表示行人图像的高度和宽度。再通过函数 F 重复应用于的局部块上，由此产生需要的条件位置编码， F 函数由内核为 $k(k \geq 3)$ 和 $\frac{k-1}{2}$ 零填充的二维卷积有效地实现。

4.1.5 拼图补丁模块

考虑到 ViT 直接的硬分割会分离出一些高度相关的区域，使用一种拼图补丁模块(JPM)，该模块能更加关注行人图像显著特征，在全局范围内探索不同局部块之间的关系，突出重要的块，抑制无用的块，进一步挖掘图片的外观信息，这些信息包含了多粒度特征和更优秀的语义信息。JPM 模块通过“移位”和“洗牌”操作对嵌入的 patch 进行重新排列，之后将它们组合成不同的部分，每个部分包含整个图像的几个随机补丁嵌入，进入共享的 Transformer 层，以端到端的方式进行训练，能够较好地缓解 ViT 中简单的 tokens 化方案所导致的背景干扰和应对数据集中行人图像错位情况的问题，将 ViT 的最后一层改为 2 个独立的分支 Global Branch 和 Jigsaw Branch，以此来学习全局特征和局部特征。

移位操作：假设输入到最后一层的隐藏特征表示为： $z_{l-1} = [z_{l-1}^0; z_{l-1}^1, z_{l-1}^2, \dots, z_{l-1}^N]$ ，将前 m 个 patch([cls]标记除外)被移到末尾也就是： $[z_{l-1}^1, z_{l-1}^2, \dots, z_{l-1}^N]$ 以 m 步移动变为 $[z_{l-1}^{m+1}, z_{l-1}^{m+2}, \dots, z_{l-1}^N, z_{l-1}^1, z_{l-1}^2, \dots, z_{l-1}^m]$ 。

洗牌操作：移位的 patch 通过 k 组的 patch 混洗操作进一步混洗，局部特征 f_1^k 可以覆盖不同部位的 Patch，JPM 通过共享 Transformer 层将其编码为 k 个局部特征 $f_1^1, f_1^2, \dots, f_1^k$ ，最终利用 k 个局部特征和全局特征 f_g 训练模型。联合使用交叉熵损失和加权正则化三元组损失^[92]。对网络中全局特征 f_g 和 k 个局部特征进行监督训练。交叉熵损失表示如式(4-2)所示：

$$L_{cls} = -\frac{1}{N} \sum_{i=1}^N \sum_{k=1}^K y_{i,k} \log(p_{i,k}) \quad (4-2)$$

加权正则化三元组损失继承了三元组损失优化正负样本之间的距离的优点，避免了引入范围参数，表示如式(4-3)所示：

$$L_{wrt}(i) = \log \left(1 + \exp \left(\sum_j w_{ij}^p d_{ij}^p - \sum_k w_{ik}^n d_{ik}^n \right) \right) \quad (4-3)$$

w_{ij}^p 和 w_{ik}^n 表示如(4-4)，(4-5)所示：

$$w_{ij}^p = \frac{\exp(d_{ij}^p)}{\sum_{d_{ij}^p \in P_i} \exp(d_{ij}^p)} \quad (4-4)$$

$$w_{ik}^n = \frac{\exp(-d_{ik}^n)}{\sum_{d_{ik}^n \in N_i} \exp(-d_{ik}^n)} \quad (4-5)$$

综上所述，本文最终损失函数如式(4-6)所示：

$$L = L_{cls}(f_g) + L_{wrt}(f_g) + \frac{1}{k} \sum_{j=1}^k (L_{cls}(f_l^j) + L_{wrt}(f_l^j)) \quad (4-6)$$

在处理过程中，将全局特征和局部特征 $[f_g, f_l^1, f_l^2, \dots, f_l^k]$ 作为最后的特征表示，更加关注身体局部的细粒度信息特征，并且拥有全局判别能力。

4.1.6 轻量化模块

ViT 自从被提出以来，因其优秀的性能被广泛的用做许多视觉任务默认的 backbone，凭借着 ViT 独特的结构，许多视觉任务的 SoTA 都得到了进一步提升，这得益于 ViT 的注意力模块，该模块存在一个 Softmax 操作，Softmax 用于在 token 上产生概率分布，这是 Transformer 的非常重要的操作之一，因为 Softmax 操作涉及到内部的计算所有输入的指数之和，所以它的计算代价相当昂贵。M 等^[93]提出将自注意力机制中的 softmax 操作替换为 ReLu 除以序列长度，缩短了模型运行时间。Han 等^[94]提出一种新的线性注意力模块，在保持模型表达能力的前提下降低了复杂度。

考虑到 ViT 对行人重识别这样的复杂场景鲁棒性不佳且训练的计算量比 CNN 要庞大。为了提高 ViT 的运行效率，缓解性能下降和模型过拟合等问题，本文设计出 DAD-T 轻量化模块缓解网络模型的负面影响。DAD-T 模块分为两部分，包括在模型中使用正则化方法 DropKey 和动态剪枝操作。

DropKey 操作目的是通过随机 drop 部分 key 的方式来自适应地调整 ViT 中的注意力权重，其变得更加平滑，它可以当作 dropout 操作的改进^[86]。DropKey 分为两种形式 DropKey-Cross 和 DropKey-Block，与 dropout 的对比如图 4-7。其中：查询(Q)，键(K)，值(V)。

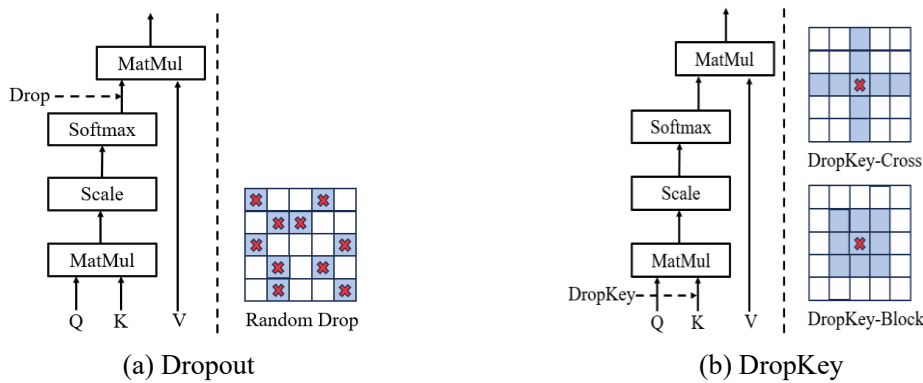


图 4-7 对比图

Fig. 4-7 Comparison chart

一般的 Dropout 是完全随机的 drop；DropKey-Cross 是以一点为中心的十字形连续区域进行 drop；DropKey-Block 是以一点为中心的矩形区域进行 drop。本文选取 DropKey-Cross 方法进行实验。

此外,为了进一步加速网络的收敛速度,缓解因为数据增强(填充、翻转、随机擦除等)、标签的处理(label smoothing 和 noise label)等技巧的混合使用而导致的内存占用、参数、运算次数、推理时间和功耗的增加的问题。考虑到当前的模型并不能在复杂的训练范式下以较少的更新次数实现对模型的快速训练,这进一步加剧了模型复杂度。使用动态剪枝采样的方法,维护梯度更新期望^[87]。

在网络训练过程时,参数、内存占用等问题会导致资源利用率增加,为了不产生冗余计算和多余开销,首先记录样本的损失值(loss)来作为样本分数,初始化默认保留首个 epoch 所有样本,之后对其余 epoch 按照剪枝概率 r 来随机对分数小于平均值的样本进行剪枝(保留概率为 $1-r$)。为了防止残余的梯度期望值偏差对结果的影响,在最后几个轮次中训练完整的数据来维持性能。

设置剪枝概率的表达式为:

$$P_t(z) = \begin{cases} r, & H_t(z) < \bar{H}_t \\ 0, & H_t(z) \geq \bar{H}_t \end{cases} \quad (4-7)$$

式中, $H_t(z)$ 和 $\bar{H}_t$ 分别表示, 样本 z 在上一轮中的分数和上一轮分数的均值。当 $H_t(z) < \bar{H}_t$ 也就是分数小于均值并留下参与训练的样本, 采用了重缩放(rescaling), 将其梯度变为 $1/(1-r)$ 。这使得整体更新是接近于无偏的。记原始数据集为 D , t 刻剪枝后的数据集为 S_t , 证明推导过程为:

经过剪枝并重缩放的操作之后, 每个样本的采样率变成 $(1-P_t(z))\rho(z)$, 缩放系数为 $\gamma_t(z) = 1/(1-P_t(z))$, 上式变为:

$$\operatorname{argmin}_{\theta \in \Theta} \mathbb{E}_{z \in S_t} [\gamma_t(z) \mathcal{L}(z, \theta)] = \operatorname{argmin}_{\theta \in \Theta} \frac{1}{c_t} \int_z \mathcal{L}(z, \theta) \rho(z) dz \quad (4-8)$$

剪枝并重缩放后的优化目标和原始的优化目标有相同的解, 因为给定时刻的 $\frac{1}{c_t}$ 是一个常数系数。

其中:

$$\mathbb{E}[\frac{1}{c_t}] = \frac{|D|}{\sum_{z \in D} (1-P_t(z))} \simeq \frac{|D|}{|S_t|} \Rightarrow \mathbb{E}[\nabla_{\theta} \mathcal{L}(S_t)] \simeq \frac{|D|}{|S_t|} \mathbb{E}[\nabla_{\theta} \mathcal{L}(D)] \quad (4-9)$$

剪枝后的更新步数变为了原先的 $\frac{|S_t|}{|D|}$, 步长变为了原先的 $\frac{|D|}{|S_t|}$, 因此概率剪枝加重缩放的策略维护了更新量的总体基本一致^[15]。

4.1.7 对比实验

本节提出的方法在 3 个主流公开数据集上分别与近年来的主流方法进行了对比。表 4-1 结果显示在这些数据集上 Rank-1 值和 mAP 值都取得了比较理想的结果。

表 4-1 不同先进方法在数据集上结果对比

Table 4-1 Comparison of results of different advanced methods on the data set

方法	DukeMTMC-ReID		MSMT17		Market-1501	
	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP
AGW ^[92]	89.0	79.6	68.3	49.3	95.1	87.8
Aligned++ ^[95]	85.2	81.2	66.3	40.7	92.8	89.4
SPM ^[96]	72.5	86.5	-	-	86.5	72.5
SCSN ^[97]	90.1	79.0	83.0	58.0	95.7	88.5
TransReID ^[46]	89.6	80.6	83.3	64.9	95.0	88.2
HAT ^[97]	90.4	81.4	82.3	61.2	95.6	89.5
PAT ^[98]	88.8	78.2	-	-	95.4	88.0
CBDBNet ^[99]	87.7	74.3	-	-	94.4	85.0
LDS ^[100]	91.5	82.5	86.5	67.2	95.8	90.4
ours	92.0	82.5	86.7	69.8	96.7	91.3

4.1.8 消融实验

本节提出的方法在一个公开数据集 Market-1501 上进行消融实验。实验结果如表 4-2 所示。

表 4-2 在数据集上的消融实验

Table 4-2 Ablation experiments on the dataset

方法	Market-1501	
	Rank-1(%)	mAP(%)
Baseline	93.6	86.0
+IceNet	94.6	88.6
+CPE	95.0	89.1
+JPM	94.3	88.0
+DAD-T	93.9	86.5
OURS	96.7	91.3

由表 4-2 可以看出，使用 Baseline 时的 Rank-1 和 mAp 为 96.7%，91.3%。在此基础上分别加上 IceNet，CPE，JPM，DAD-T 四个模块，Rank-1 指标分别达到了 94.6%，95.0%，94.3%，93.9%；mAp 指标分别达到了 88.6%，89.1%，88.0%，

86.5%。由此看出单独添加 IceNet, CPE, JPM, DAD-T 四个模块均能提高行人重识别的性能。四种方法联合使用效果最好, 最终的 Rank-1 为 96.7%, mAp 为 91.3%。证明了本文提出的方法在遮挡场景下可以有效提高行人重识别的识别精度。

4.1.9 可视化热力图

图 4-8 展示了添加本文提出模块前后热力图可视化图像对比, 热力图中暖色表示网络对行人图像某一区域的关注度, 颜色越明显, 代表关注程度越高。图像从左到右分别展示了原始行人图像、ViT 输出和本文方法输出的热力图。

结果显示本文添加模块可以很好地提升网络性能, 能够同时关注行人图像中粗粒度的全局信息和细粒度的局部信息, 验证了本模型出色的鲁棒性能。

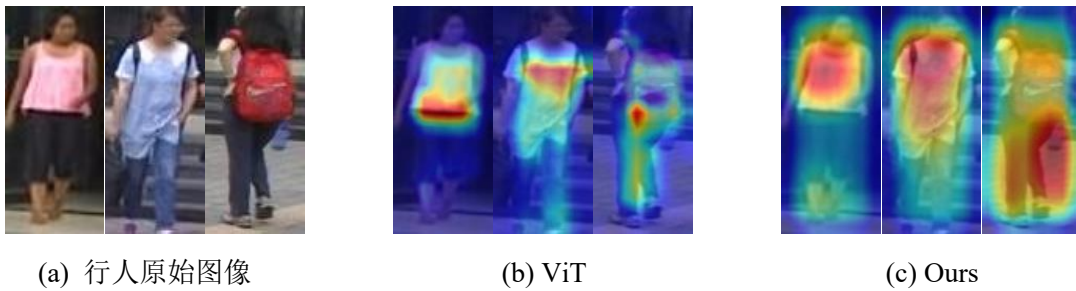


图 4-8 可视化对比

Fig. 4-8 Visual comparison

4.2 本章小结

行人重识别是图像分类的子类任务, 近年来随着深度学习的发展, 这项技术被广泛应用于公共安全领域。但由于换衣、光照、遮挡等因素的干扰会导致行人重识别的准确率降低。同时由于 Transformer 计算复杂度高导致训练过程缓慢, 提出一种基于暗光增强和细粒度特征提取的行人重识别模型。首先对行人图像进行图像增强, 使行人图像更加清晰, 色彩更加明亮, 减少色差影响, 其次在 Transformer 中嵌入条件位置编码和拼图补丁模块使网络拥有平移不变性, 拥有综合提取行人全局特征和局部特征的能力。最后考虑到因为计算复杂度高导致训练过程缓慢的问题, 为了加速网络的收敛速度, 添加了轻量化模块, 提高了模型的鲁棒性和泛化能力。实验在三个行人重识别公开数据集上进行训练和评价。结果显示本文提出方法有效缓解了暗光场景下行人重识别准确率低的问题。

第 5 章 总结与展望

5.1 工作总结

本文首先说明了当前行人重识别的研究背景和研究意义,然后详细介绍了行人重识别在有监督领域和无监督领域的研究现状。之后介绍了行人重识别任务遇到的各类研究难点和挑战,常用的大型公共数据集以及性能评价标准。

近年来,随着深度学习的不断发展,越来越多的先进方法应用到图像视觉领域。传统的卷积神经网络模型虽然具有较好的泛化性能,能够对大量样本进行反复迭代计算,并结合特征学习,从而提升其识别精度,但由于行人重识别任务较为复杂,对其进行特征提取具有很大的挑战。在此背景下,本文的主要研究成果如下:

(1)鉴于卷积神经网络的感受野存在一定的局限性,难以有效地聚焦于关键特征的提取,同时有监督的行人重识别方法通常过度依赖大量标注数据,从而限制了其在实际场景中的广泛应用。基于这些考虑,本文提出了一种基于无监督学习的行人重识别方法。首先,通过对数据集中的行人图像进行随机遮挡操作来模拟真实世界中的遮挡情况,同时利用对比裁剪技术来减轻遮挡和复杂背景对神经网络的干扰。接着,引入了一种全新的多尺度特征提取网络 SC-CAResNet,并将其与对比学习相结合,以提高对行人图像特征的捕获能力。最后,采用多种损失函数联合训练网络。研究结果表明,该方法显著提升了模型在遮挡场景下行人重识别任务的性能,为解决现实场景中的行人重识别难题提供了有效的解决方案。

(2)Transformer 模型近年来备受关注,其出色的网络结构具有全局感受野,能够有效提取全局特征。相对于传统的卷积神经网络,Transformer 能更好地捕捉图像中远距离像素之间的依赖关系。本文将条件位置编码、拼图补丁和 DAD-T 模块引入 Vision Transformer 模型中,并在暗光环境下进行实验验证。研究结果显示,这种新方法能够有效增强网络对暗光环境下行人图像的细粒度特征信息的关注,提升行人特征的提取能力,同时降低模型复杂度,缓解性能下降问题。这一研究为在暗光环境下行人重识别任务提供了一种有效的解决方案,有望在实际应用中取得更好的性能表现。

5.2 未来展望

行人重识别任务作为跨摄像头目标检索的关键技术,在智慧城市安防,疫情防控溯源,零售行为分析等方面发挥重要作用。但在复杂现实场景下面临多重技术瓶颈,具有诸多挑战和机遇。

传统方法在处理光照变化、遮挡和复杂背景等现实场景下的问题时存在局限性,比如:在光照剧烈变化场景中易出现特征漂移(如夜间红外监控与白天可见

光场景的模态差距);严重遮挡(如雨伞、交通工具遮挡)导致基于全局特征匹配的方法失效,现有局部对齐算法对非刚性形变的建模能力不足;密集人群中的姿态突变、快速运动模糊等动态干扰,造成时空关联特征提取困难;大规模跨摄像头标注成本高昂,现有监督学习模型在无标注目标域(如新建监控系统)中性能急剧退化。

本文提出的结合随机遮挡与对比裁剪的无监督行人重识别方法在遮挡场景下虽然取得良好的性能,然而从技术实现与应用落地角度审视,仍存在多方面改进空间。模拟遮挡场景与真实场景遮挡物形态存在一定差距,改进后的模型虽增强了特征提取能力,但计算复杂度提升,在小微型设备部署时可能面临效率问题,同时跨层级特征融合机制有待优化以强化浅层特征遮挡鲁棒性。

基于暗光增强和细粒度特征提取的行人重识别方法提高了识别精度,但在增强策略的语义导向性、模型效率优化、多模态场景适配及工程化评估等方面仍有提升空间。后续研究可围绕自适应光照增强、轻量化时空建模、多模态特征融合及多维度评估体系展开,推动行人重识别技术向复杂真实环境下的精准化、高效化应用迈进。

针对上述问题,新一代深度学习框架展现出突破性潜力,为解决这些问题提供了新的可能性。这些技术有效推动了行人重识别在实际应用中取得更大的突破和进展,为构建更智能、安全的社会做出了贡献。

参考文献

- [1]王素玉, 肖塞. 行人重识别研究综述[J]. 北京工业大学学报, 2022, 48(10): 1100-1112.
- [2]Zhang B, Shao X, Chen W, et al. Visual tracking based on robust appearance model[J]. Image and Vision Computing, 2019, 89: 211-221.
- [3]Zheng Z, Zheng L, Yang Y. A discriminatively learned cnn embedding for person reidentification[J]. ACM transactions on multimedia computing, communications, and applications (TOMM), 2017, 14(1): 1-20.
- [4]Sun Y, Zheng L, Yang Y, et al. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline) [C]. Proceedings of the European conference on computer vision (ECCV).2018:480-496.
- [5]Fu Y, Wei Y, Zhou Y, et al. Horizontal pyramid matching for person re-identification[C]. Proceedings of the AAAI conference on artificial intelligence. 2019, 33(01): 8295-8302.
- [6]Hermans A, Beyer L, Leibe B. In defense of the triplet loss for person re-identification[J]. arXiv preprint arXiv:1703.07737, 2017.
- [7]Lv X, Zhao C, Chen W. A novel hard mining center-triplet loss for person re-identification[C]. Chinese Conference on Pattern Recognition and Computer Vision (PRCV). Cham: Springer International Publishing, 2019: 199-210.
- [8]Chen W, Chen X, Zhang J, et al. Beyond triplet loss: a deep quadruplet network for person re-identification[C]. Proceedings of the IEEE conference on computer vision and pattern recognition, 2017:403-412.
- [9]Hadsell R, Chopra S, Lecun Y. Dimensionality reduction by learning an invariant mapping[C]. 2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06). IEEE, 2006, 2: 1735-1742.
- [10]Chopra S, Hadsell R, Lecun Y. Learning a similarity metric discriminatively, with application to face verification[C]. 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05). IEEE, 2005, 1: 539-546.
- [11]田青, 王斌, 周子梹. 无监督行人重识别研究综述[J]. 计算机工程, 2024: 1-22.
- [12]Yang W, Zhang D. Unsupervised person re-identification by part-compensated soft multi-label learning[J]. IET Image Processing, 2022, 16(7): 2012-2024.
- [13]Wang D, Zhang S. Unsupervised person re-identification via multi-label classification[C]. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 10981-10990.

- [14]Zhai Y, Lu S, Ye Q, et al. Ad-cluster: Augmented discriminative clustering for domain adaptive person re-identification[C]. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 9021-9030.
- [15]Zou Y, Yang X, Yu Z, et al. Joint disentangling and adaptation for cross-domain person re-identification[C]. Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16. Springer International Publishing, 2020: 87-104.
- [16]Wang M, Lai B, Huang J, et al. Camera-aware proxies for unsupervised person re-identification[C]. Proceedings of the AAAI conference on artificial intelligence. 2021, 35(4): 2764-2772.
- [17]Jia X, Wang X, Mi Q. An unsupervised person re-identification approach based on cross-view distribution alignment[J]. IET Image Processing, 2021, 15(11): 2693-2704.
- [18]Lee G, Lee S, Kim D, et al. Camera-driven representation learning for unsupervised domain adaptive person re-identification[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023: 11453-11462.
- [19]Chen H, Wang Y, Lagadec B, et al. Joint generative and contrastive learning for unsupervised person re-identification[C]. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 2004-2013.
- [20]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, 2020, 33: 11309-11321.
- [21]Zhang L, Song R, Wang Y, et al. Dual-GCL: Dual-Graph Contrastive Learning for Unsupervised Person Re-Identification[J]. IEEE Transactions on Cognitive and Developmental Systems, 2024, 16(4): 1352-1363.
- [22]Luo X, Jiang M, Kong J, et al. Hierarchical Camera-Aware Contrast Extension for Unsupervised Person Re-Identification[J]. IEEE Transactions on Multimedia, 2024.
- [23]Zeng K, Ning M, Wang Y, et al. Hierarchical clustering with hard-batch triplet loss for person re-identification[C]. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 13657-13665.
- [24]Yang F, Zhong Z, Luo Z, et al. Joint noise-tolerant learning and meta camera shift adaptation for unsupervised person re-identification[C]. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 4855-4864.
- [25]Li H, Hu Q, Hu Z. Catalyst for Clustering-Based Unsupervised Object Re-identification: Feature Calibration[C]. Proceedings of the AAAI Conference on Artificial Intelligence.2024, 38(4): 3091-3099.

- [26]Xuan S, Zhang S. Intra-inter camera similarity for unsupervised person re-identification[C]. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 11926-11935.
- [27]Zou C, Chen Z, Cui Z, et al. Discrepant and multi-instance proxies for unsupervised person re-identification[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023: 11058-11068.
- [28]Zhang X, Li D, Wang Z, et al. Implicit sample extension for unsupervised person re-identification[C]. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 7369-7378.
- [29]Cho Y, Kim W J, Hong S, et al. Part-based pseudo label refinement for unsupervised person re-identification[C]. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 7308-7318.
- [30]Si T, He F, Wu H, et al. Spatial-driven features based on image dependencies for person re-identification[J]. Pattern Recogn, 2022, 124: 108462.
- [31]Si T, He F, Zhang Z, et al. Hybrid contrastive learning for unsupervised person re-identification[J]. IEEE Transactions on Multimedia, 2022, 25: 4323-4334.
- [32]Wang X, Liu M, Wang F, et al. Relation-preserving feature embedding for unsupervised person re-identification[J]. IEEE Transactions on Multimedia, 2023, 26: 714-723.
- [33]张敏, 余增, 韩云星, 等. 面向复杂场景的行人重识别综述[J]. 计算机科学, 2022, 49(10): 138-150.
- [34]Zheng L, Shen L, Tian L, et al. Scalable person re-identification: A benchmark[C]. Proceedings of the IEEE international conference on computer vision. 2015: 1116-1124.
- [35]Ristani E, Solera F, Zou R, et al. Performance measures and a data set for multi-target, multi-camera tracking[C]. European conference on computer vision. Cham: Springer International Publishing, 2016: 17-35.
- [36]Li W, Zhao R, Wang X. Human reidentification with transferred metric learning [C]. Asian conference on computer vision. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012: 31-44.
- [37]Wei L, Zhang S, Gao W, et al. Person transfer gan to bridge domain gap for person re-identification[C]. Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 79-88.
- [38]Lin T-Y, Maire M, Belongie S, et al. Microsoft coco: Common objects in context[C]. Computer vision - ECCV 2014: 13th European conference, zurich,

Switzerland, September 6-12, 2014, proceedings, part v 13. Springer International Publishing, 2014: 740-755.

[39]Zheng L, Bie Z, Sun Y, et al. Mars: A video benchmark for large-scale person re-identification[C]. Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI 14. Springer International Publishing, 2016: 868-884.

[40]He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]. Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.

[41]Huang W, Luo M, Zhang P, et al. Full-scaled deep metric learning for pedestrian re-identification[J]. Multimedia Tools and Applications, 2021, 80: 5945-5975.

[42]Liu Z, Mao H, Wu C-Y, et al. A convnet for the 2020s[C]. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 11976-11986.

[43]Liu Z, Lin Y, Cao Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows[C]. Proceedings of the IEEE/CVF international conference on computer vision. 2021: 10012-10022.

[44]Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in Neural Information Processing Systems, 2017, 30.

[45]Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[J]. preprint arXiv:2010.11929, 2020.

[46]He S, Luo H, Wang P, et al. Transreid: Transformer-based object re-identification [C]. Proceedings of the IEEE/CVF international conference on computer vision. 2021: 15013-15022.

[47]Li J, Wang M, Gong X. Transformer based multi-grained features for unsupervised person re-identification[C]. Proceedings of the IEEE/CVF winter conference on applications of computer vision. 2023: 42-50.

[48]王静, 李沛潼, 赵容锋, 等. 融合卷积注意力和 Transformer 架构的行人重识别方法[J]. 北京航空航天大学学报, 2024, 50(02): 466-476.

[49]赵倩, 薛超晨, 赵琰. 基于改进型 Transformer 编码器和特征融合的行人重识别[J]. 数据采集与处理, 2023, 38(02): 375-385.

[50]Zhu K, Guo H, Zhang S, et al. AAformer: Auto-Aligned Transformer for Person Re-Identification[J]. IEEE Transactions on Neural Networks and Learning Systems, 2023.

- [51]Tay C-P, Roy S, Yap K-H. Aanet: Attribute attention network for person re-identifications[C]. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 7134-7143.
- [52]Song C, Huang Y, Ouyang W, et al. Mask-guided contrastive attention model for person re-identification[C]. Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 1179-1188.
- [53]Zhuo J, Chen Z, Lai J, et al. Occluded person re-identification[C]. 2018 IEEE international conference on multimedia and expo (ICME). IEEE, 2018: 1-6.
- [54]何磊, 栗风永, 秦川. 跨通道交互注意力机制驱动的双流网络跨模态行人重识别[J]. 应用科学学报, 2024, 42(05): 884-892.
- [55]Dai Z, Wang G, Yuan W, et al. Cluster contrast for unsupervised person re-identification[C]. Proceedings of the Asian conference on computer vision. 2022: 1142-1160.
- [56]Wu Z, Xiong Y, Yu S X, et al. Unsupervised feature learning via non-parametric instance discrimination[C]. Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 3733-3742.
- [57]He K, Fan H, Wu Y, et al. Momentum contrast for unsupervised visual representation learning[C]. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 9729-9738.
- [58]Grill J-B, Strub F, Althé F, et al. Bootstrap your own latent-a new approach to self-supervised learning[J]. Advances in neural information processing systems, 2020, 33: 21271-21284.
- [59]Oord A V D, Li Y, Vinyals O. Representation learning with contrastive predictive coding[J]. arXiv preprint arXiv:1807.03748, 2018.
- [60]Tian Y, Krishnan D, Isola P. Contrastive multiview coding [C]. Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16. Springer International Publishing, 2020: 776-794.
- [61]Chen T, Kornblith S, Norouzi M, et al. A simple framework for contrastive learning of visual representations[C]. International conference on machine learning. PmLR, 2020: 1597-1607.
- [62]Wang G, Yuan Y, Chen X, et al. Learning discriminative features with multiple granularities for person re-identification [C]. Proceedings of the 26th ACM international conference on Multimedia. 2018: 274-282.

- [63]Yan G, Wang Z, Geng S, et al. Part-Based Representation Enhancement for Occluded Person Re-Identification[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2023, 33(8): 4217-4231.
- [64]Wang T, Liu H, Song P, et al. Pose-guided feature disentangling for occluded person re-identification based on transformer[C]. Proceedings of the AAAI conference on artificial intelligence. 2022, 36(3): 2540-2549.
- [65]库向阳, 李蕊心, 叶鸥. 融合随机擦除和残差注意力网络的行人重识别[J]. 计算机工程与应用, 2022, 58(03): 215-221.
- [66]Woo S, Park J, Lee J-Y, et al. Cbam: Convolutional block attention module[C]. Proceedings of the European conference on computer vision (ECCV). 2018: 3-19.
- [67]南方, 陈国钊, 朱艺璇, 等. 基于软掩膜前景分割和多信息融合重排序的行人重识别方法[J]. 光学技术, 2023, 49(06): 756-762.
- [68]汪旭, 胡晓光, 付哲宇, 等. 基于 Transformer 的全局-局部融合特征的遮挡行人重识别方法[J]. 计算机科学与探索, 2024: 1-22.
- [69]周玉, 赵小锋, 汪一, 等. 关键细粒度信息指导的多尺度遮挡行人重识别[J]. 电子与信息学报, 2024, 46(06): 2578-2586.
- [70]Devries T, Taylor G W. Improved Regularization of Convolutional Neural Networks with Cutout[J]. arXiv preprint arXiv:1708.04552, 2017.
- [71]Peng X, Wang K, Zhu Z, et al. Crafting better contrastive views for siamese representation learning[C]. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 16031-16040.
- [72]Chen H, Lagadec B, Bremond F. Ice: Inter-instance contrastive encoding for unsupervised person re-identification[C]. Proceedings of the IEEE/CVF international conference on computer vision. 2021: 14960-14969.
- [73]Ester M, Kriegel H-P, Sander J, et al. A density-based algorithm for discovering clusters in large spatial databases with noise[C]. kdd. 1996, 96(34): 226-231.
- [74]Li J, Wen Y, He L. Scconv: Spatial and channel reconstruction convolution for feature redundancy[C]. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023: 6153-6162.
- [75]Hou Q, Zhou D, Feng J. Coordinate attention for efficient mobile network design[C]. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 13713-13722.
- [76]Hu J, Shen L, Sun G. Squeeze-and-excitation networks[C]. Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 7132-7141.

- [77]Kiran M, Praveen R G, Nguyen-Meidine L T, et al. Holistic guidance for occluded person re-identification[J]. arXiv preprint arXiv:2104.06524, 2021.
- [78]Tian Q, Du X. Multi-class center dynamic contrastive learning for unsupervised domain adaptation person re-identification[J]. Computers and Electrical Engineering, 2024, 116: 109155.
- [79]Ye P, Zeng H, Zhang W, et al. Part-aware network: a simple but efficient method for occluded person re-identification[C]. International Conference on Computer Application and Information Security (ICCAIS 2021). SPIE, 2022, 12260: 158-166.
- [80]Li Y, Zhu X, Sun J, et al. Unsupervised person re-identification based on high-quality pseudo labels[J]. Applied Intelligence, 2023, 53(12): 15112-15126.
- [81]Kim M, Cho M, Lee H, et al. Occluded person re-identification via relational adaptive feature correction learning [C]. ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2022: 2719-2723.
- [82]Tan H, Liu X, Yin B, et al. MHSA-Net: Multihead self-attention network for occluded person re-identification[J]. IEEE Transactions on Neural Networks and Learning Systems, 2022, 34(11): 8210-8224.
- [83]Luo H, Gu Y, Liao X, et al. Bag of tricks and a strong baseline for deep person re-identification[C]. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops. 2019: 0-0.
- [84]Ko K, Kim C S. IceNet for Interactive Contrast Enhancement[J]. Ieee Access, 2021, 9: 168342-168354.
- [85]Chu X, Tian Z, Zhang B, et al. Conditional positional encodings for vision transformers[J]. arXiv preprint arXiv:2102.10882, 2021.
- [86]Li B, Hu Y, Nie X, et al. Dropkey for vision transformer[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 22700-22709.
- [87]Qin Z, Wang K, Zheng Z, et al. Infobatch: Lossless training speed up by unbiased dynamic data pruning[J]. arXiv preprint arXiv:2303.04947, 2023.
- [88]Luo W, Li Y, Urtasun R, et al. Understanding the effective receptive field in deep convolutional neural networks[J]. Advances in neural information processing systems, 2016, 29.
- [89]Wang X, Girshick R, Gupta A, et al. Non-local neural networks[C]. Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 7794-7803.

- [90]周传华, 周东东, 夏徐东, 等. 基于卷积注意力机制和多损失联合的跨模态行人重识别[J]. 微电子学与计算机, 2022, 39(06): 22-30.
- [91]曾涛, 薛峰, 杨添. 面向行人重识别的通道与空间双重注意力网络[J]. 计算机工程, 2022, 48(12): 281-287+295.
- [92]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, 2021, 44(6): 2872-2893.
- [93]Wortsman M, Lee J, Gilmer J, et al. Replacing softmax with relu in vision transformers[J]. arXiv preprint arXiv:2309.08586, 2023.
- [94]Han D, Pan X, Han Y, et al. Flatten transformer: Vision transformer using focused linear attention[C]. Proceedings of the IEEE/CVF international conference on computer vision. 2023: 5961-5971.
- [95]Luo H, Jiang W, Zhang X, et al. AlignedReID plus plus : Dynamically matching local information for person re-identification[J]. Pattern Recogn, 2019, 94: 53-61.
- [96]Zhang S, Zhang Q, Yang Y, et al. Person re-identification in aerial imagery[J]. IEEE Transactions on Multimedia, 2020, 23: 281-291.
- [97]Chen X, Fu C, Zhao Y, et al. Saliency-guided cascaded suppression network for person re-identification[C]. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 3300-3310.
- [98]Li Y, He J, Zhang T, et al. Diverse part discovery: Occluded person re-identification with part-aware transformer[C]. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 2898-2907.
- [99]Tan H, Liu X, Bian Y, et al. Incomplete descriptor mining with elastic loss for person re-identification[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2021, 32(1): 160-171.
- [100]Zang X, Li G, Gao W, et al. Learning to disentangle scenes for person re-identification[J]. Image and Vision Computing, 2021, 116: 104330.

攻读学位期间发表的学术成果

已录用的论文:

[1] 第一作者, 基于暗光增强和细粒度特征提取的行人重识别[J]. 陕西理工大学学报(自然科学版)(已录用)

[2] 第一作者, Unsupervised person re-identification with random occlusion and ContrastiveCrop [C]. The 14th International Conference on Brain-Inspired Cognitive Systems(已录用)

致谢

时光荏苒，三年的学习生活即将画上句号。回想起这段旅程，心中满是感激。在完成这篇硕士论文的漫长旅程里，我得到了太多人的帮助。

我要先谢谢我的导师。导师严谨的治学态度，一直激励着我，让我不敢有丝毫懈怠，是导师让我明白了做学问该有的态度。

实验室的师兄姐妹们，也给了我莫大的支持。大家一起在图书馆查阅资料，在实验室讨论研究方案，分享各自的经验，分析问题。

还有我的家人，始终是我坚实的后盾。感谢他们毫无保留的爱，让我能安心完成学业。

回首这段时光，满是成长与收获。感恩一路上所有的遇见！

