分类号_____ 密级 _____

UDC _____

学 位 论 文

基于特征融合的行人重识别方法的研究与实现

作 者 姓 名 ：

指 导 教 师 ：

东北大学计算机科学与工程学院

申请学位级别： 硕士 学 科 类 别 ： 专业学位

学科专业名称： 计算机技术

论文提交日期： 年 月 论文答辩日期： 年 月

学位授予日期： 年 月 答辩委员会席：

评 阅 人 ：

东 北 大 学

年 月

摘 要

行人重识别（Person Re-identification）是一个图像检索的子问题，其目的是从行人图像集合中选取一个待查询图像(Query)，然后从跨设备采集的行人图像库（Gallery）中查找与 Query 相同身份的行人图像。比如在疫情当下，需要对确诊病例进行流行病学调查时，便可以利用行人重识别技术快速地从来自多个监控摄像头的行人图像集合中，检索出属于病例的图像样本。从而确定其活动轨迹，以便在最短时间内进行相应的防疫部署。行人重识别研究中仍有许多需要攻克的难点，本文进行了如下工作：

（1）提出了基于多分辨特征与空间信息融合的行人重识别模型。通常行人重识别采用预训练的卷积神经网络作为特征提取骨干网络，但是卷积过程存在对浅层特征中信息丢失问题。并且，现有行人重识别数据集图像中往往存在无用的背景干扰，这对于行人重识别算法的精度有着非常大的影响。针对如上问题，本文对骨干网络进行了两方面改进：其一，提出多分辨率特征融合的行人重识别方法，通过提取模型不同层次的卷积特征进行融合，并且引入通道注意力机制保证融合过程中不引入浅层干扰信息，增强模型利用来自不同层次卷积特征中的信息的能力。其二，提出空间信息融合的行人重识别方法。通过在骨干网络中引入空间注意力机制，并且在全局特征提取分支引入全局对比池化，使模型在特征提取阶段可以对行人和背景进行判别，以减少背景信息的干扰。最后，在 Market-1501 和 DukeMTMC-reID 数据集上进行了实验，并通过 rank-1 和 mAP 指标验证了本文提出的方法对模型识别精度的提升。

（2）提出了基于局部特征融合的切块对齐行人重识别模型。提取细粒度的局部特征是目前行人重识别方法的一个主流研究方向。基于预定义切块的方法是一种低成本且高效的局部特征提取方法。但是，现有数据集中存在空间中不对齐行人图像，基于切块的方法在处理此类图像时会引入不对齐误差，从而降低性能指标。针对这一问题，本文提出了如下两方面改进：一方面，针对不对齐问题对行人重识别数据集中的图像进行了细粒度分类，并以此对模型误判过程进行了详细的分析。然后在局部分支根据切块间相似度，对原始样本在局部切块特征层面进行相似度度量与融合对齐，以减少图像不对齐问题引入的局部距离度量误差。另一方面，通过双分支信息融合的困难样本挖掘方法，在不引入过多计算量的情况下，使局部分

支在模型训练过程中获得更多会引起不对齐问题的样本。最后，选择上述同样的数据集和实验指标设置了实验以及详细的可视化工作，证明了算法对模型性能的提升，并对算法原理进行了分析。

（3）最后基于本文提出的模型，进行了行人重识别可视化系统的设计与实现，用于可视化本文中的行人识别结果。

关键词：行人重识别；深度学习；特征融合；局部特征

Abstract

Person Re-identification is a sub-problem of image retrieval, which aims to select a query image (Query) from a collection of pedestrian images, and then find pedestrian images with the same identity as the Query from a cross-device collection of pedestrian images (Gallery). For example, in the case of an epidemiological investigation of a confirmed case in an epidemic situation, Person Re-identification can be used to quickly retrieve an image sample belonging to a case from a collection of pedestrian images from multiple surveillance cameras. This allows the trajectory to be determined so that the corresponding epidemic prevention can be deployed in the shortest possible time. There are still many difficult points to be overcome in Person Re-identification research, and the following work is conducted in this thesis.

(1) A Person Re-identification model based on the fusion of multi-resolution features and spatial information is proposed. Usually, pedestrian re-recognition uses a pre-trained convolutional neural network as the feature extraction backbone network, but the convolution process has the problem of information loss in shallow features. Moreover, there is often useless background interference in the existing pedestrian re-recognition dataset images, which has a great impact on the accuracy of Person Re-identification algorithm. To address the above problems, this thesis improves the backbone network in two aspects: first, it proposes a pedestrian re-recognition method with multi-resolution feature fusion, by extracting the convolutional features of different levels of the model for fusion, and introduces a channel attention mechanism to ensure that no shallow interference information is introduced in the fusion process, and increases the utilization of information in the model for different levels of features. Second, the Person Re-identification method with spatial information fusion is proposed. By introducing the spatial attention mechanism in the backbone network and introducing global contrast pooling in the global feature extraction branch, the model can discriminate between pedestrians and background in the feature extraction stage to reduce the interference of background information. And experiments are conducted on Market-1501 and DukeMTMC-reID datasets, and the proposed method and optimization strategy are

verified to be effective in improving model accuracy by rank-1 and mAP metrics.

(2) The slice aligned network is proposed. Extracting fine-grained local features is a mainstream research direction of Person Re-identification methods at present. The method based on predefined cut blocks is a low-cost and efficient local feature extraction method. However, there are unaligned pedestrian images in space in the existing dataset, and the block-based method introduces unalignment errors when processing such images, which degrades the performance index. To address this problem, two improvements are proposed in this thesis: on the one hand, a fine-grained classification of the images in the pedestrian re-recognition dataset is performed to address the misalignment problem, and the model misclassification process is analyzed in detail in this way. Then, local alignment calculation is performed on the original samples at the local feature level according to the inter-slice similarity in the local branch to reduce the distance metric error introduced by image misalignment. On the other hand, the difficult sample mining method with two-branch information fusion enables the local feature branch to obtain more samples that cause the misalignment problem during model training without introducing too much computational effort. Finally, the same dataset and the same experimental metrics mentioned above are chosen to set up experiments as well as detailed visualization work to validate the principle performance enhancement of the algorithm and perform a rationale analysis.

(3) Finally, based on the model proposed in this thesis, the design and implementation of a Person Re-identification visualization system are carried out for visualizing the algorithm results in this thesis.

Key words: person re-identification; deep learning; feature fusion; local features

目 录

第 1 章 绪论

1.1 研究背景与意义

随着社会和科技发展，越来越多新兴技术逐渐融入并服务于我们的日常生活，计算机视觉相关技术就是其中之一。在公共安防领域，随着技术的进步，摄像头已经成为了价格低廉且应用广泛的监控设备。我国的天网工程，就是利用安装在城市公共区域，特别是公共聚集场所的监控摄像头，进行图像、视频采集，并对这部分数据加以利用，保障城市公共安全。

如今视频采集随着摄像头安装的边际效应，硬件成本已经可以控制得很低。但在实际的跨摄像机行人检索的场景中，往往还是通过如图 1.1 中人工的方式进行筛查。这一过程需要消耗大量的人力和时间成本。而且在海量的监控视频的面前，仅以人工的方式无法同时满足识别准确性和时效性的要求。

图 1.1 人工方式进行行人检索

Fig 1.1 pedestrian retrieval in manual mode

如图 1.2 所示，行人重新识别（Person-ReID）的任务是在不同的时间、不同的摄像机下的行人图像集合中，给出一个特定行人并进行检索的问题。这一任务的主要挑战来自于包括照明情况不同、摄像机参数不一致以及拍摄行人角度变化引起的姿态不一致等问题[1]。

对于行人重识别领域的学术研究，最开始是来自于另一个热门的研究领域，跨摄像头多目标跟踪 MTMC（Multi-target Multi-camera Tracking）问题[1]。跨摄像头目标跟踪需要在多个摄像头的视野下，对同一个目标进行轨迹追踪。论文[3]中利用目标相似度，来进行多摄像头情况下的轨迹关联。其中重中之重，便是如何度量目标之间的相似度。论文[4]，则系统地对行人重识别的研究目标和边界进行了界定，明确了行人重识别方法的研究应当着力于行人的特征提取和相似度度

量。而后续在行人重识别领域的研究工作也都主要着力于研究如何提取更具有判别性的行人特征，保证经过相似度度量后得到更好的识别结果。

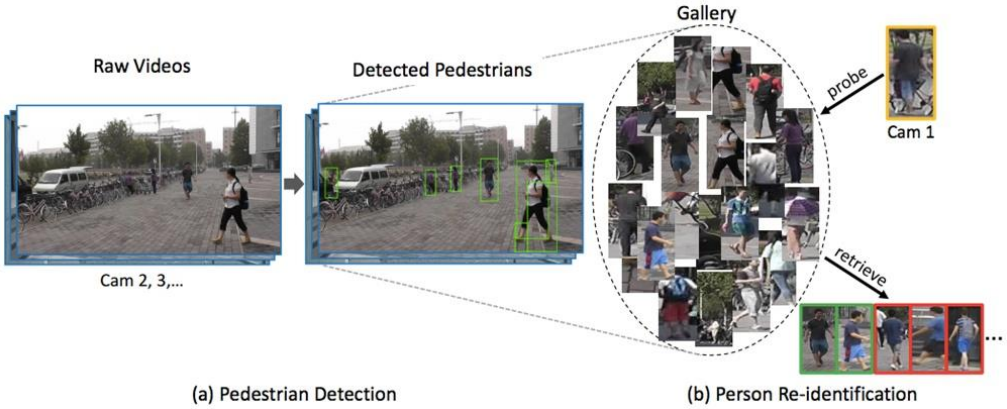


图 1.2 行人检测与行人重识别流程图[4]

Fig 1.2 flow chart of pedestrian detection and pedestrian re recognition[4]

因此，本文旨在研究如何提取出更具有判别性的特征，这是一个极具挑战性，且具有实际应用意义的课题。并且随着行人重识别领域相关研究，将计算机视觉技术应用在服务于民生大众日常生产生活场景中，提升城市安全，发展智慧城市、智慧安防已经是当前的大趋势。因此对行人重识别方法研究作为其中重要的一环具有重要的意义。

1.2 国内外研究现状

正如上文所描述，行人重识别的主要研究目标是检索具有来自不同摄像头的相同身份的行人，其过程主要包括特征提取和相似度度量两个步骤。传统的个人身份识别方法需要手动识别标记人员目标，这会消耗大量人力成本。传统的方法中使用的手工图像特征有颜色、HOG（Histogram of oriented gradient）[6]、SIFT（Scale invariant feature transform）[7]等。然后，利用 XQDA（Cross-view Quadratic Discriminant Analysis）[8]或者 KISSME（Keep It Simple and Straightforward Metric Learning）[9]来学习最佳的相似度度量。随着在深度神经网络的发展，近年来出现了许多基于深度学习的行人重识别方法。

本节总结了近年来国内外发表的行人重识别研究方法，对基于深度学习的行人重识别方法进行了分类。并主要从四个方面介绍近年来国内外行人重识别领域的研究现状，包括局部特征学习、基于全局特征与局部特征融合学习、基于生成对抗网络和基于序列特征学习的方法。

1.2.1 基于局部特征的方法

行人重识别方法可以从特征提取的粒度进行分类[1]。最为常见的是对整张图像进行粗粒度的整体特征提取，即基于全局特征的行人重识别方法[10][12]。这类方法在特征提取时关注整个图像空间平面内的内容，但是对于行人细节特征捕捉能力不强。因此，随着行人重识别相关研究的发展，出现了一系列提取细粒度局部特征的方法，用于捕捉图像中行人细节信息。

基于局部特征的方法通常需要对行人图像进行局部划分并确保每个局部特征的对齐，再通过神经网络对过这些区域进行提取特征。

常用的局部特征提取方法有基于预定义的切块分割的方法[13],[14]、基于多尺度融合的方法[20]-[21]、基于软注意的方法[22]-[24]、基于行人语义提取[25]-[28]和基于全局和局部特征联合的方法[29],[30]。这些方法通常以解决遮挡、边界检测错误、视图和姿势变化等问题为出发点，并开展相关研究。

本节将选取其中部分类型中具有代表性的方法进行介绍。

（1）基于切块的局部特征提取方法

Sun[15]等人考虑了每个切块内的内容一致性，提出了一个基于局部特征的基线网络 PCB。PCB 采用垂直方向上均匀划分策略来学习局部特征，输出由多个切块组成的卷积特征。并且提出了 RPP 方法以增强每个区块内的特征内容的一致性，从而保证切块在空间上对齐。

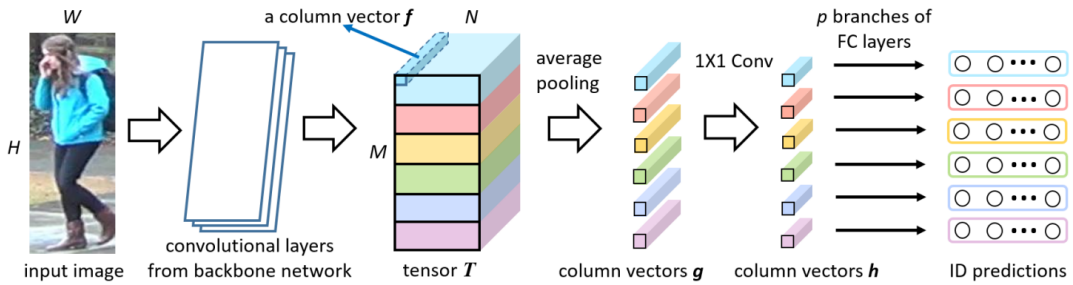


图 1.3 基于切块的局部特征提取网络 PCB[15]

Fig 1.3 block based local feature extraction network PCB[15]

为了解决在特征提取阶段图像中杂乱的背景信息和行人的服饰引入的干扰信息，Fu[14]通过提出水平金字塔池化 HPP 的结构。HPP 主要的思路可以概括为两个方面：一是通过不同尺度的滑框，即文中的水平金字塔尺度，对行人图像进行不同粒度的局部特征进行提取，使后续的特征中包含不同层级的细节信息；二是文章中提出全局最大值池化可以更好提取前景信息，全局平均值池化可以更好

提取整体信息，但是相对应地包含更多背景信息。通过融合的方式可以利用两者优势，减弱背景的干扰。

（2）基于网格的局部特征提取方法

Sun [16]提出了可见性感知零件模型 VPM，用来解决行人图像空间不对齐和行人图像对之间语义信息不对等的问题。VPM 使模型具有自动判断行人图像完整性的能力，并且图像存在缺失和完整性不同的情况下，可以进行不同程度的分割。该研究在一定程度上减少了因为行人局部语义缺失和空间平面内人体拓扑结构不对齐引入的误差。

（3）基于多尺度特征融合的方法

不同尺度的特征包含语义信息是不同的，提取多尺度的行人特征进行融合，可以获得丰富的行人特征表示。由于在不同尺度学习的行人特征存在差异或冲突，多尺度特征直接融合可能无法达到最佳融合效果。因此，研究多尺度特征在融合过程中如何进行优势互补也是极为重要的[2]。

在行人重识别相关研究中，基于卷积神经网络的方法在进行图像输入时通常需要进行图像处理。在这一过程中要保证模型最终的输出维度统一，需要将行人数据集中的图像进行调整，将输入的图像尺寸放缩使其输入模型时是一致的。如果图像原尺寸是大于统一输入尺寸的，在这一过程中会被裁剪或进行放缩，这样的操作无疑都会丢失原始图像中的细节内容。所以 Chen [33]提出一种深度金字塔特征学习网络框架 DPFL，能够学习多尺度的互补特征，并且克服跨尺度特征学习时的差异问题，可以从既有的行人图像中挖掘出更多可用信息。

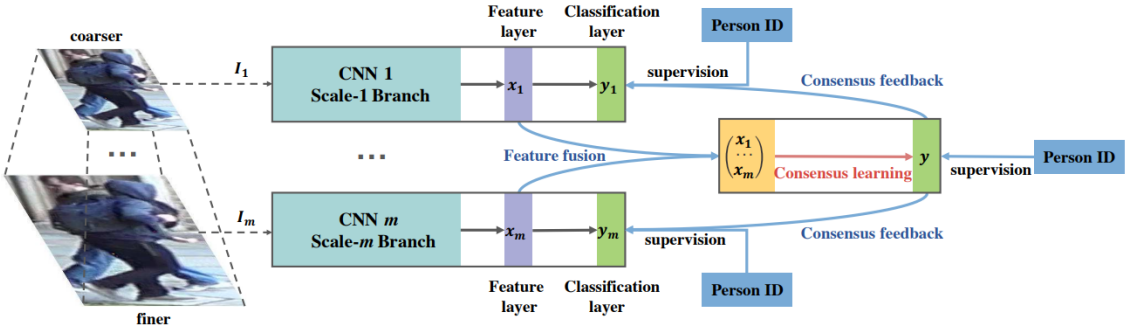


图 1.4 DPFL 模型结构图[33]

Fig 1.4 DPFL model structure[33]

（4）基于软注意力的方法

注意力相关方法的目标是找到对特征图有更大影响的区域，并使模型能够集中在身体外观等有区别的局部部位上，以纠正错位和消除背景干扰[35]。注意力机制在计算机视觉领域的良好表现，也经常被用于行人重识别任务中的基于局部特

征的相关方法中。目前大多数基于注意力的行人重识别方法倾向于使用软注意力，可分为空间注意力、通道注意力、混合注意力、非局部注意力和位置注意力[2]。

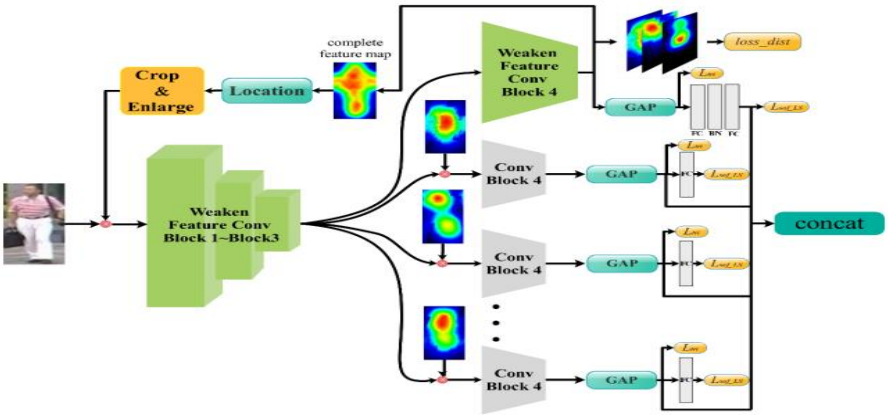


图 1.5 多分支注意力模型[36]

Fig 1.5 multi branch attention model[36]

为了获得行人的细粒度局部特征，Ning [36]提出了一种多分支注意力网络，通过去除干扰信息的自适应滤波获得局部特征。在行人重新识别任务中，用注意力机制提取细粒度信息的方法已在大量方法中被证明是有效的。然而，提取的细粒度信息仍可能包括冗余信息。此外，目前的方法缺乏去除背景干扰的有效方案。因此，这篇研究中提出了特征细化和过滤网络方法，从三个方面解决上述问题：首先，通过弱化高响应特征，识别高价值特征并提取人的完整特征，从而增强模型的鲁棒性；其次，模型通过定位和截取人物的高响应区域，消除背景信息的干扰，加强模型对行人完整特征的响应；最后，使用多分支注意力网络选择最有价值的细粒度特征进行行人重识别，以提高模型的性能。

（5）基于姿态模型的方法

另外一个热门研究方向通过引入其他深度神经网络来提取身体部位或姿势等语义信息。这一类方法能够有效地提高局部特征提取准确率，从而提升行人重识别模型的性能。

Zhao [25]等人将人体拓扑结构信息在行人重识别任务中加以应用，并提出了一种新的网络结构 SpindleNet。首先，SpindleNet 使用姿态估计模型定位身体部位的 14 个关键点，提取出行人的 7 个身体区域。然后，使用卷积神经网络从上述不同的身体区域捕获语义特征。最后，合并来自不同身体区域的语义特征。SpindleNet 可以对齐身体部位的特征，并且可以更好地突出局部细节信息。

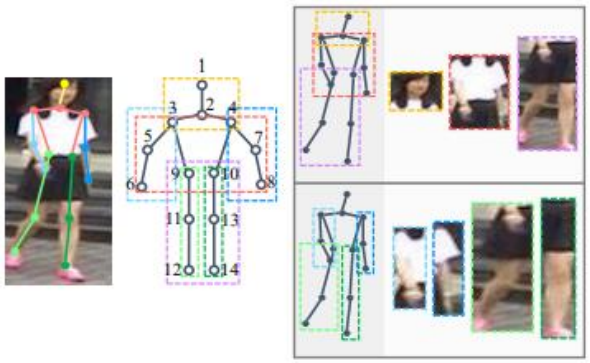


图 1.6 基于姿态模型的方法 SpindleNet[25]

Fig 1.6 SpindleNet method based on attitude model[25]

1.2.2 基于全局与局部特征融合的方法

局部特征学习可以捕获有关行人区域的详细信息，但局部特征的可靠性可能会受到姿势和遮挡变化的影响。因此，一些研究经常将细粒度的局部特征与粗粒度的全局特征结合起来，以增强最终的特征表示[29]。

Ming[72]设计了一种将全局和局部特征动态对齐的模型 GLDFA-Net，从分利用了全局特征和局部特征各自的优势。传统的基于局部的方法主要集中在学习行人预定义区域的局部特征。这些方法通常使用局部对齐方法或引入关键人体姿态点等辅助信息来匹配局部特征，这在遇到较大场景差异时往往不适用。GLDFANet 在局部分支中引入了局部滑动对齐 LSA 策略来指导距离度量的计算，从而提高测试阶段的准确性。

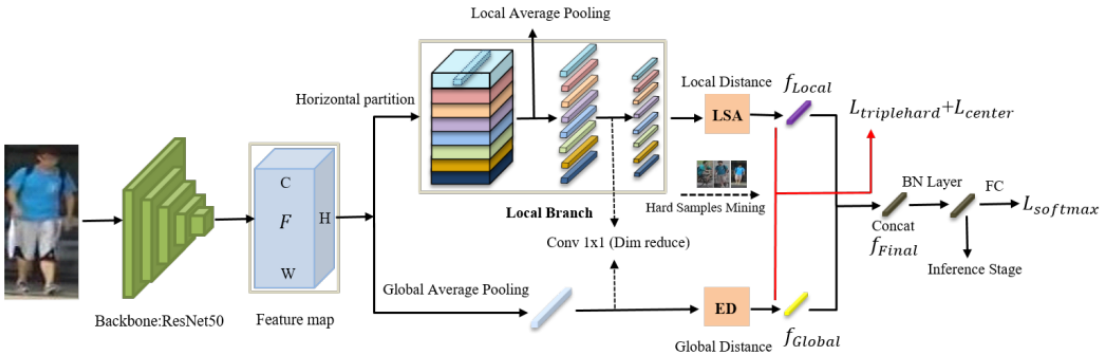


图 1.7 GLDFA-Net 模型结构[72]

Fig 1.7 GLDFA-Net model structure [72]

1.2.3 基于生成对抗网络的方法

生成对抗网络 GAN[5]一般用于图像合成，但是今年来一些研究将此方法引入行人重识别的研究领域之中。行人重识别图像按照视觉内容差异进行分类，可

以分为两个类别。一种差异是以分辨率差异、明度差异和色像差异为主。此类差异称之为低级差异。令一种则是来自图像语义层面，以行人姿态差异、摄像头视角差异和遮挡物差异为主。称之为高级差异[2]。

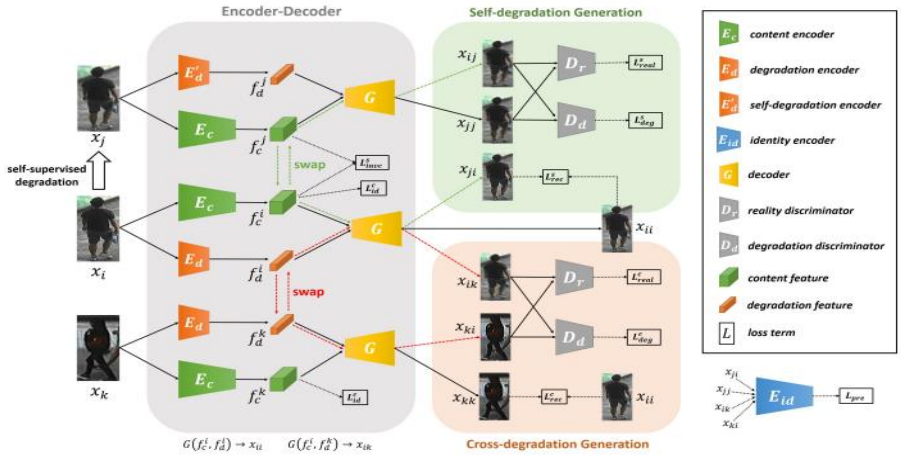


图 1.8 DDGAN 模型结构[38]

Fig 1.8 ddgan model structure[38]

为了解决上述问题，特别是解决低级差异问题，Huang[38]等人提出了一种可以用于真实世界行人重识别的退化不变性学习框架。通过引入自监督学习策略，能够在没有额外监督的情况下消除图像退化。并且该框架只需进行少量修改，就可以很容易地扩展到解决其他的退化因素的问题。

1.2.4 基于序列特征的方法

基于序列特征的行人重识别方法也是研究的热点。这些基于序列特征的方法以短视频为输入，同时使用空间和时间作为互补，可以缓解单一基于外观特征的局限性[2]。这些方法中的大多数使用光流信息、三维卷积神经网络、递归神经网络或长短期记忆、时空注意方法或图卷积网络对视频序列的时空信息进行建模[2]。这些方法可以减轻遮挡、分辨率变化、光照变化、视图和姿势变化等因素对识别结果的影响。

（1）基于光流的方法

在基于视频序列的行人重识别方法中，将序列视为沿时间维度进行排列的行人图像帧。利用光流法可以通过像素变化获得时间维度上各帧之间的关系，以获得行人的运动信息[39]。

Kipf [40]提出了一种累积上下文网络 AMOC，它由两个输入序列组成，其中分别输入原始 RGB 图像和包含运动信息的光流图像。AMOC 通过光流信息来提

高最终模型识别的准确性。

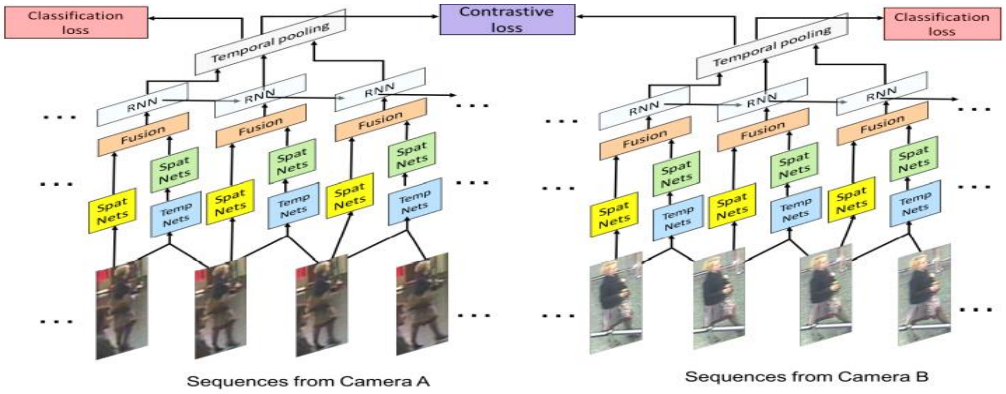


图 1.9 AMOC 模型结构[40]

Fig 1.9 AMOC model structure [40]

（2）基于图卷积的方法

近年来，图卷积网络 GCN[41]由于其强大的能力而被广泛用于行人重识别任务，并且出现了大量基于此结构的变体网络[42]-[46]。

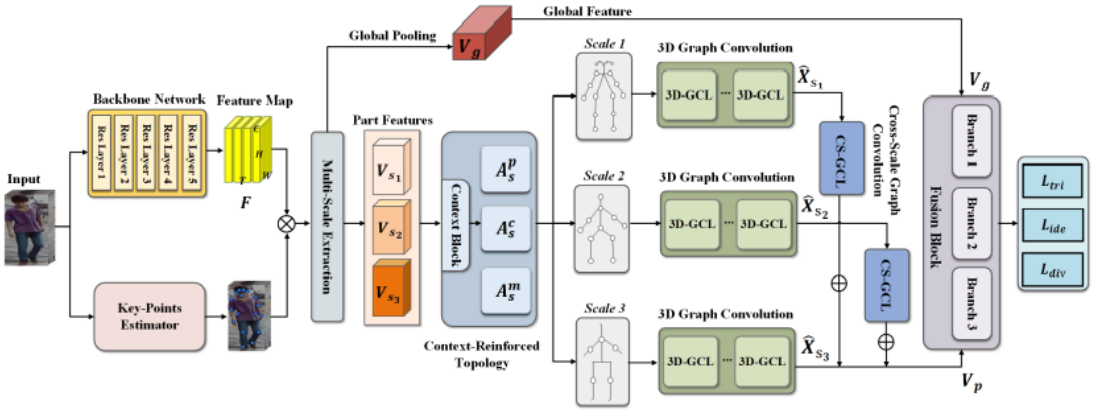


图 1.10 CTL 模型结构[47]

Fig 1.10 CTLmodel structure[47]

基 GCN 的方法性能明显优于其他序列特征学习方法。特别是 Liu[47]等人提出的 CTL 方法，利用卷积神经网络主干和关键点估计器从多个粒度的人体中提取语义局部特征作为图节点，有效地挖掘与外观信息互补的综合信息，增强最终特征的表征能力。

1.3 主要工作和内容安排

1.3.1 主要工作

（1）基于多分辨特征与空间信息融合的行人重识别模型

首先，结合近年来的研究提出一种基于全局特征融合的行人重识别网络结构。

随后，通过引入目前基于深度学习的行人重识别方法中用到的模型策略进行工程优化提升模型性能，得到一个强有力的基线。然后基于现有基线结构中可优化的部分，提出一种基于多分辨率特征融合的行人重识别方法，以解决在骨干网络在深层卷积过程中对底层信息的忽略问题。最后，并引入空间信息融合方法，解决图像背景造成的样本全局特征杂糅无用背景信息的问题。

（2）基于局部特征融合的切块对齐行人重识别模型

提出一种基于局部特征融合的切块对齐方法，用于解决行人图像在空间垂直方向拓扑结构不对齐的问题，以增强模型的判别能力，获得更好的行人重识别性能。然后，在局部分支对图像进行预定义方式的水平切块获取，并通过样本间的局部距离矩阵获取切块间的注意力矩阵。最后，通过使用切块间的注意力对原始样本特征进行融合，以获得垂直方向对齐的局部特征描述符，以减小行人垂直方向拓扑结构不对齐带来的误差。

（3）行人重识别可视化系统设计与实现

结合前文研究的得到的算法模型，对行人重识别可视化系统的进行分析并针对最小需求集合进行编码实现。

1.3.2 内容安排

第一章，绪论部分。首先，本章主要从技术和社会发展角度，结合场景需求，介绍了行人重识别研究的背景和意义。然后，介绍了行人重识别领域目前国内外的研究现状。针对本文的关注点，着重介绍了基于局部特征和基于全局特征的一些行人重识别方法，并对前沿的基于生成对抗网络和基于视频序列的行人重识别也做了介绍。最后，对本文的主要研究内容和文章结构安排进行的描述。

第二章，相关技术介绍。本文主要针对基于深度学习的行人重识别方法进行研究，所以在这一章中将所涉及的一部分深度学习相关基础研究进行介绍。首先，将介绍卷积神经网络相关的基础。然后针对本文后续将使用的骨干网络进行介绍，包括基本的残差网络结构和 ResNet 基本结构介绍。最后，将介绍一些今年的基于深度学习的行人重识别方法中，对于模型训练时使用到的能够提升模型性能的训练策略进行了汇总和介绍。

第三章，基于多分辨特征与空间信息融合的行人重识别模型。首先，结合近年来的研究提出一种基于全局特征融合的行人重识别网络结构，并进行优化得到一个强有力的基线。然后，针对现有的网络结构，结合卷积神经网络的特点，描

述了模型在终层对底层特征信息缺失的问题，并提出了基于多分辨率特征融合的思路。接着，介绍了行人重识别中常见的样本背景复杂的问题，并结合当前的网络结构，提出引入了空间信息融合模块模块以解决背景信息干扰的问题。最后，对提出的改进进行了实现，并在开源数据集上进行实验，以验证改进的结构的有效性。

第四章，基于局部特征融合的切块对齐行人重识别模型。首先，介绍了目前流行的行人重识别数据集和现实场景中，行人图像在垂直方向上拓扑结构不对齐的问题。然后针对不对齐问题对模型引入干扰问题进行了分析，并提出了解决思路。最后，提出基于局部特征融合的切块对齐行人重识别模型，并在开源数据集上进行实验以验证算法的有效性，并通过可视化工作对算法原理进行了分析。

第五章，行人重识别可视化系统设计与实现。基于前文的方法研究，已经能够得到具有较高识别率的行人重识别算法和基于 Pytorch 实现的网络模型，但是直接操作框架使用模型的过程繁琐且容易出错，对使用者不友好。本章通过结合目前主流技术与上文研究成果，对行人重识别可视化系统进行了设计与实现。

第六章，总结和展望。对本文工作进行了总结，并对下一步工作方向进行了展望。

1.4 本章小结

本章一开始主要从技术和社会发展角度，结合场景需求，介绍了行人重识别研究的背景和意义。然后，针对本文的关注点介绍了行人重识别领域目前国内外的研究现状。最后，对本文的主要研究内容和文章结构安排进行了描述。

第 2 章 相关技术概述

2.1 卷积神经网络

基于神经网络的深度学习算法成为机器学习的热门分支。深度学习算法试图通过使用具有复杂结构的多个处理层来对数据进行高级抽象，并已经能够在计算机视觉、自然语言处理、机器人技术等领域的许多任务中胜过此前最先进的方法[48]。在计算机视觉领域，卷积神经网络是其中被应用最为广泛的[49]。本文中的研究基于卷积神经网络，所以首先对其进行介绍。

对于计算机视觉任务，输入数据是大小通常在数百到数万像素之间的图像。如果一个神经网络只用全连接的神经元处理这个输入矩阵，那么要训练的参数数量会非常大，过拟合的风险很高。卷积神经网络通过引入具有稀疏链接和参数共享特点的卷积层，对这一结构进行了优化。

卷积神经网络接收图像矩阵作为输入，隐藏层的卷积核仅连接到输入矩阵中的一小部分区域，所以输入和输出的空间相关性是局部的。这个区域被称为卷积核的局部感受野。此外，因为卷积核的信息抽取能力并不特定局限于图像中的某些区域，在同一层的每个卷积计算平面内可以共享相同的权重。这两个属性大大减少了网络的参数数量。通过应用这两个原理，将一个全连接的神经元层转换为一个卷积层，如下所示：

$$y = a(W * X + b) \tag{2.1}$$

对于二维图像X作为输入，卷积算子定义为：

$$(W * X)(i,j) = \sum_m \sum_n X(m,n)W(i-m, j-n) \tag{2.2}$$

其中X是输入图像W是权重矩阵，也称为过滤器或内核。b是偏置项。函数a是激活函数，运算符$*$表示离散卷积运算。

在图像处理中，卷积运算可以用于边缘检测、图像锐化和模糊，只需使用不同的滤波器矩阵的数值即可。这意味着不同的过滤器可以从图像中检测不同的特征。在卷积层中，卷积核权重矩阵W的参数，由反向传播算法自动学习。在卷积神经网络中一层通常包含多个这样的卷积核，然后将每个卷积核的输出特征图在通道维度进行叠加组成这一层最终的输出。

在卷积神经网络结构整体中，图像通过一系列卷积层、非线性层、池化层和

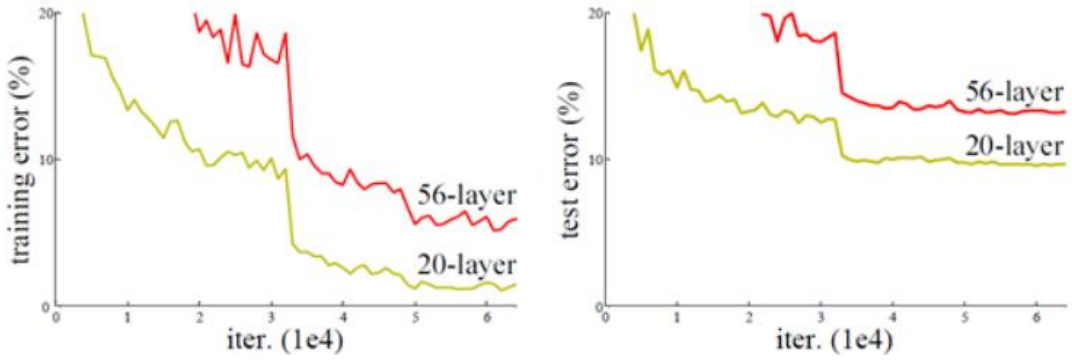
全连接层，然后生成输出。直观的角度来看，浅层的卷积操作趋向于识别图像上的边界线条和简单的颜色。提取网络深层更高级别的属性，例如行人重识别任务中的行人轮廓和姿态等，则需要更大的感受有野。

池化层一般紧随非线性层之后，常见的有最大值池化核平均值池化。池化操作对卷积层生成的特征进行下采样，降低在空间维度的大小。在这个过程中，特征中的信息会被压缩。这意味着如果在之前的卷积操作中已经识别出一些特征，例如边界核轮廓等将被压缩并提取更加高阶的语义信息。

通常，基于深度学习的行人重识别研究利用 ImageNet 预训练的卷积神经网络模型在作为骨干，并针对其自身特点进行改造，并在后续行人重识别相关的数据集上进行微调以训练模型。目前行人重识别领域使用比较广泛的卷积神经网络是 ResNet50，所以后续将对残差网络和 ResNet50 进行介绍。

2.2 残差网络

基于深度学习的网络模型，一般而言模型深度越深则进行信息提取和抽象的能力就越强。卷积神经网络也是深度神经网络的其中一种，增加模型深度在一定程度上的确能够提升在目标任务的性能指标。有研究发现，并不能够无限制地通过增加深度的方式增强模型的能力。因为，模型增加到一定深度后，模型变得更为复杂，但是模型的性能会到达瓶颈甚至是下降[56]。



（a） （b）

图 2.1 20 层与 56 层卷积神经网络模型在 CIFAR-10 上的误差[56]

Fig 2.1 error of layer 20 and layer 56 convolutional neural network models on cifar-10[56]

如图 2.1 为 20 层与 56 层卷积神经网络在 CIFAR-10 上的分类误差对比图。如图所示当模型深度由 20 增加到 56 之后，无论是（a）中训练还是（b）中推理阶段误差均明显的增加。

（1）残差学习

残差学习的提出，就是为了解决由于随网络深度增加，模型性能被抑制的问

题。如图 2.2 残差学习提出一种“短路结构”，让模型能够后在训练过程中可以形成短路结构以降低模型的复杂度。一般情况下，当输入x经过模型之后我们希望学习到信息$H(x)$。而在残差单元中，希望模型通过学习残差$F(x)=H(x)-x$进而间接得到信息$H(x)$。这样当残差$F(x)$为 0 值时，在此层输入到输出只是做了恒等变换，相当于在模型结构上进行了剪枝操作。

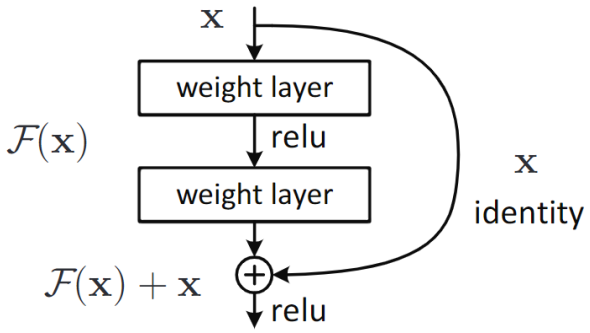


图 2.2 残差单元[56]

Fig 2.2 residual element[56]

基于以上思想，将网络中l层残差单元输入极为x_l，输出记为x_{l+1}，则可以得到如下计算式：

$$y_l = h(x_l) + F(x_l, W_l) \tag{2.3}$$

$$x_{l+1} = f(y_l) \tag{2.4}$$

其中h为恒等变换，F表示残差函数，f为$ReLU$函数。那么该层的学习特征为：

$$X_L = x_l + \sum_{i=l}^{L-1} F(x_i, W_i) \tag{2.5}$$

则可以求得x_l处梯度为：

$$\frac{\partial loss}{\partial x_l} = \frac{\partial loss}{\partial x_L} \cdot \frac{\partial x_L}{\partial x_l} \tag{2.6}$$

那么，由(2-5)可得：

$$\frac{\partial x_L}{\partial x_l} = \frac{\partial}{\partial x_l}\left(x_l + \sum_{i=l}^{L-1} F(x_i, W_i)\right) = 1 + \frac{\partial}{\partial x_l}\sum_{i=l}^{L-1} F(x_i, W_i) \tag{2.7}$$

最后，由(2-6)和(2-7)可得：

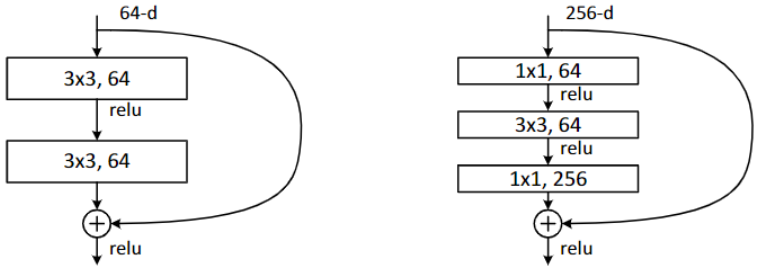
$$\frac{\partial loss}{\partial x_l} = \frac{\partial loss}{\partial x_L}\left(1 + \frac{\partial}{\partial x_l}\sum_{i=l}^{L-1} F(x_i, W_i)\right) \tag{2.8}$$

上式表明当损失反向传播到 L 层残差块时，若计算结果为 1，则表示短路，此时梯度仍能继续传递。并且残差总是一个比较小的值，所以模型训练更加容易。

（2）ResNet

ResNet 就是将残差学习的思想，引入到卷积神经网络的一种具体实现。如图

2.3 为卷积神经网络中的残差模块结构图。

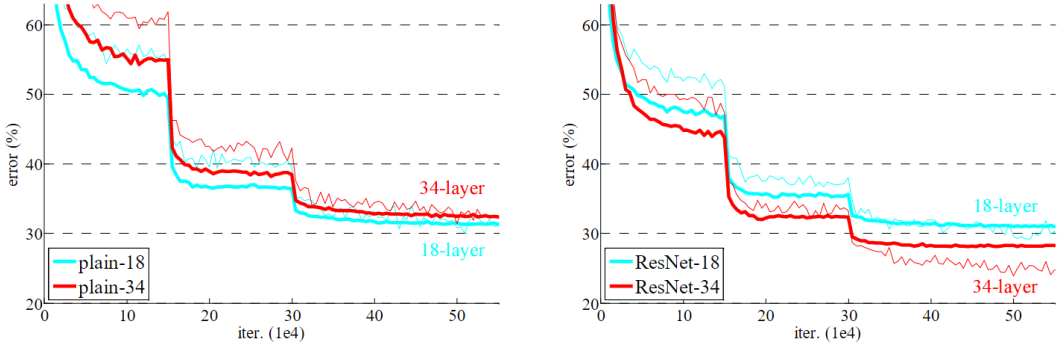


（a） （b）

图 2.3 卷积神经网络中的残差单元[56]

Fig 2.3 residual element in convolutional neural network[56]

ResNet 将残差引入间距神经网络并在 ImageNet 进行了相关实验。如图 2.4 可见为 ResNet18 和 ResNet34 在 ImageNet 上的实验效果。图中(a)为普通卷积神经网络结构，相同条件下 18 层的网络性能由于 34 层的网络性能。图中(b)显示，加入残差单元后，当网络深度由 18 增加至 34，具有更深层次的 ResNet34 取得的成绩明显由于了前者。



（a） （b）

图 2.4 不同深度的模型在 ImageNet 效果对比[56]

Fig 2.4 comparison of models with different depths in Imagenet[56]

layer name	output size	18-layer	34-layer	50-layer	101-layer	152-layer
conv1	112×112	7×7, 64, stride 2				
conv2_x	56×56	3×3 max pool, stride 2				
		$\begin{bmatrix}3\times3,64\\3\times3,64\end{bmatrix}\times2$	$\begin{bmatrix}3\times3,64\\3\times3,64\end{bmatrix}\times3$	$\begin{bmatrix}1\times1,64\\3\times3,64\\1\times1,256\end{bmatrix}\times3$	$\begin{bmatrix}1\times1,64\\3\times3,64\\1\times1,256\end{bmatrix}\times3$	$\begin{bmatrix}1\times1,64\\3\times3,64\\1\times1,256\end{bmatrix}\times3$
conv3_x	28×28	$\begin{bmatrix}3\times3,128\\3\times3,128\end{bmatrix}\times2$	$\begin{bmatrix}3\times3,128\\3\times3,128\end{bmatrix}\times4$	$\begin{bmatrix}1\times1,128\\3\times3,128\\1\times1,512\end{bmatrix}\times4$	$\begin{bmatrix}1\times1,128\\3\times3,128\\1\times1,512\end{bmatrix}\times4$	$\begin{bmatrix}1\times1,128\\3\times3,128\\1\times1,512\end{bmatrix}\times8$
conv4_x	14×14	$\begin{bmatrix}3\times3,256\\3\times3,256\end{bmatrix}\times2$	$\begin{bmatrix}3\times3,256\\3\times3,256\end{bmatrix}\times6$	$\begin{bmatrix}1\times1,256\\3\times3,256\\1\times1,1024\end{bmatrix}\times6$	$\begin{bmatrix}1\times1,256\\3\times3,256\\1\times1,1024\end{bmatrix}\times23$	$\begin{bmatrix}1\times1,256\\3\times3,256\\1\times1,1024\end{bmatrix}\times36$
conv5_x	7×7	$\begin{bmatrix}3\times3,512\\3\times3,512\end{bmatrix}\times2$	$\begin{bmatrix}3\times3,512\\3\times3,512\end{bmatrix}\times3$	$\begin{bmatrix}1\times1,512\\3\times3,512\\1\times1,2048\end{bmatrix}\times3$	$\begin{bmatrix}1\times1,512\\3\times3,512\\1\times1,2048\end{bmatrix}\times3$	$\begin{bmatrix}1\times1,512\\3\times3,512\\1\times1,2048\end{bmatrix}\times3$
	1×1	average pool, 1000-d fc, softmax				
FLOPs		1.8×10^9	3.6×10^9	3.8×10^9	7.6×10^9	11.3×10^9

图 2.5 不同深度的 ResNet 模型结构细节[56]

Fig 2.5 structural details of RESNET model at different depths[56]

由于残差学习，可以使模型通过短路的方式自行对复杂度进行调整，演化出而来层次结构越来越深的网络结构。如图为深度不同的 ResNet 模型结构细节。

目前行人重识别领域比较常用的为 ResNet50，或者基于 ResNet50 的优化结构，本文所研究的方法也基于此进行。

2.3 行人重识别模型训练策略

本文中所涉及的行人重识别方法是基于深度学习进行研究的，深度神经网络模型的性能和效果往往和数据以及训练策略密不可分。通过设置更好的训练策略，可以在不影响原有模型算法的情况下，提升最终识别性能。本节整理了目前基于深度学习的行人重识别方法，在模型训练过程中涉及到的一些用于提升模型性能的策略。

2.3.1 预热学习率

学习率的设置在深度神经网络模型的训练过程中至关重要，正确的学习率设置策略往往会对模型性能带来巨大提升。行人重识别方法中通常会采用衰减策略，即在训练初期设置较大的学习率，使模型能够大致拟合样本分布，然后使用较小的学习率，使损失稳定在极小值附近[50]。

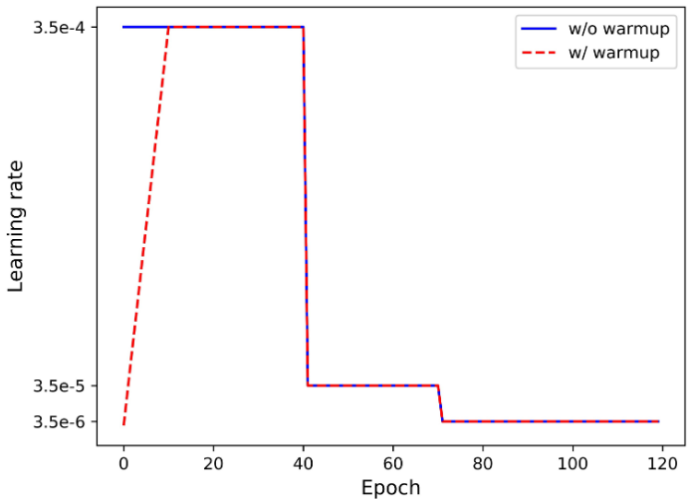


图 2.6 预热学习率策略[50]

Fig 2.6 preheating learning rate strategy[50]

而行人重识别网络通常会采用与训练模型初始化作为骨干网络，由于行人重识别的数据集和与训练数据集分布不一致，初期如果选择较大的学习率将会使与训练模型的性能下降[51]。所以如图 2.6 基于预热的学习率，在行人重识别领域的

方法中被广泛应用，并且在正确设置学习率的情况下能够明显提升模型性能。

2.3.2 标签平滑损失

行人重识别方法中重要的一类，是将行人判别看作是行人分类任务进行模型训练，分类损失使用交叉熵损失。模型在训练过程中，将会尽力去拟合 one-hot 标签中的 0 值和 1 值。当数据中出现少量样本标签错误的情况，将会出现极大的损失值，导致模型训练被干扰[53]。使用标签平滑的损失，模型在训练过程中会在一定程度上去匹配目标值而不是尝试全量匹配。标签平滑，可以有效地减少错误标签带来的影响，在一定程度上还能减小模型的过拟合[50]。

训练集中记存在 N 张训练样本，其属于 K 个行人，并且我们知道其对应标签$y(i) = [y_1, y_2, \dots, y_k]$。其中$y(i)$是行人标签的 one-hot 表示，当样本属于某一行人时，对应位数值为 1，否则记为 0。

在模型末端分类层，类别输出$z = [z_1, z_2, \dots, z_k]$,则我们可知该样本为第 i 个行人的概率为:

$$p(i) = \frac{\exp(z_i)}{\sum_{k=1}^{K} \exp(z_k)} \tag{2.9}$$

则分类损失为：

$$L_{cls}(x) = -\sum_{k=1}^{K} y_k log p(k) \tag{2.10}$$

使用标签平滑可以使模型减小对标签值的过度依赖，具体做法如式 3-4 所示，给标签添加一个极小错误率ε，使模型在一定程度上增大类别匹配位置的数值，减少其他位置处的数值，而非极力去使预测值接近绝对的 0 值和 1 值。

$$y'_i = \begin{cases} \varepsilon & , y_i = 0 \\ 1 - \varepsilon, & y_i = 1 \end{cases} \tag{2.11}$$

考虑上行人重识别数据集本身的类别数，采用如式 2.12 所示进行标签平滑计算：

$$q_i = \begin{cases} \dfrac{\varepsilon}{M} & , y_i = 0 \\ 1 - \dfrac{M-1}{M}\varepsilon, & y_i = 1 \end{cases} \tag{2.12}$$

其中，M为数据集中的行人个数，由交叉熵损失函数可得标签平滑损失如式 2.13 所示，最终网络模型训练的损失L'_{cls}如下：

$$L'_{cls}(x) = -\sum_{i=1}^{K} q_i log p(i) \tag{2.13}$$

2.3.3 多损失联合策略

近年来的行人重识别方法中，通常都会使用多种损失函数对模型进行优化。其中最为常见的是使用分类损失和三元组损失组合进行组合，对模型进行联合训练。在。前文中介绍过，重识别的主要流程可以抽象为特征提取和相似度度量两个过程。传统只是用分类损失进行训练的模型中，会在推理阶段将最终的分类层的神经元进行丢弃，使用剩下的模型结构对行人图像进行特征提取，然后再进行距离度量。通过此类损失训练的模型，更容易收敛，但是最终模型适合使用余弦相似度进行度量。而随着行人重识别数据集的行人类别增多，发现三元组损失更适合用于行人重识别模型。而三元组损失优化目的是将行人图像在特征空间了分为不同的簇，实现类内紧凑而类间区隔，欧式距离更加适合作为距离度量。如果简单将两种损失进行相加确定的损失，在训练时容易出现震荡，不利于训练[50]。

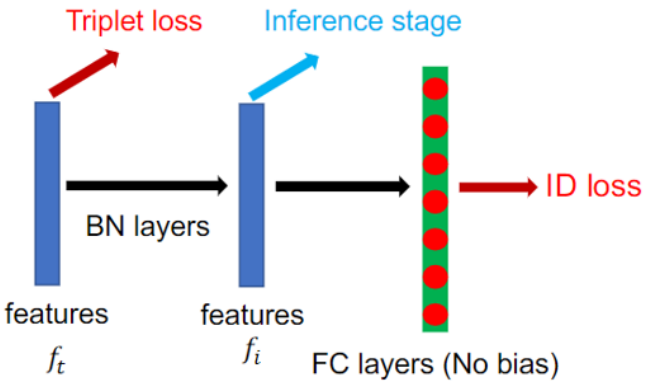


图 2.7 BN 层增加位置[50]

Fig 2.7 added position of BN layer[50]

为了解决此类问题，可以在模型分类层之前添加批量归一化 BN（Batch Normalization）[54]层，并进行初始化。如图 2.7 为 BN 层在模型中添加的位置。f_t为原始结构中用于训练三元组损失和分类损失的特征，通过 BN 层得到f_i特征。使用f_t来训练三元组损失，f_i进行分类损失，可以使模型训练更加平稳。

2.3.4 数据增广

深度神经网络随着模型深度的增加，对原始数据信息提取和建模的能力随之增强，但是参数量的增加也带来了过拟合的风险[48]。解决这一问题的方法有很多，可以通过正则化和早停等方式抑制模型过渡拟合。另外，直接进行数据增广，使模型在训练过程中输入更接近于真实分布的数据，对防止过拟合也是十分有效的

[52]。

行人重识别任务中同一个目标的不用图像样本，由于来自多摄像头，受其视角、距离和光照等因素影响，其图像内容也大为不同。所以会在训练过程中对原始图像进行随机擦除、裁切、旋转和明暗变化，以降低模型进行判别使对遮挡、行人大小、偏角和亮度的敏感性。

2.4 本章小结

本章针对本文后续会使用到的相关技术进行了汇总和介绍，包括卷积神经网络、残差网络和行人重识别模型训练策略。

第 3 章 基于多分辨特征与空间信息融合的行人重识别模型

3.1 研究动机

近年来，行人重识别领域涌现出了许多高精度的识别方法。例如，Luo[57]提出 Alignedreid 在行人重识别问题上超过了人类表现，并提供了一个采用全局特征和局部特征共同学习的基线。此外，论文[50]使用简单的全局特征，并通过优化模型训练策略，获得了更高的识别准确率。上诉方法中的模型，均直接使用了 ResNet50 作为骨干网络进行特征提取，所以无法充分利用浅层卷积特征。并且仅使用单一池化直接计算全局特征，容易受背景信息干扰。

根据前文介绍，可以对行人重识别任务的关键步骤进行概括，主要分为特征提取和距离度量两个阶段。在特征提取阶段提取出更具有判别性的特征，最终行人重识别方法的性能就更好。在计算机视觉领域，卷积神经网络是提取图像特征常用的工具。在卷积神经网络中，随着对图像进行逐层卷积操作，浅层卷积特征也逐渐随着池化操作和更深层的卷积操作汇聚成深层的高级语义特征。如颜色和边缘轮廓等浅层信息汇聚成行人姿态信息。在这一过程中前者可能出现丢失，但这一些信息在行人识别的过程中极有可能作为比后者重要的判别依据[12]。

此外行人图像经过卷积网络得到三维特征描述符，需要再通过全局池化得到一维特征向量作为行人的整体特征。通常方法中这一步会使用最大值池化或平均值池化。但是在求最值的同时会舍弃其他特征值，也就丢弃了包含在其中的一些有用判别信息。平均值池化在面对背景占比高的图像时会引入过多无用信息[73]。

针对如上问题本文提出双分支行人重识别模型，并加入了多分辨率特征融合方法增强模型对不同层次特征的利用能力，随后加入空间信息融合方法使模型能够对行人和背景进行不同程度的关注，增强模型抵抗图像中背景信息干扰的能力。具体而言，首先将模型各个卷积阶段中不同分辨率的特征进行提取并融合，将不同层次卷积特征中的信息进行汇总。接着，在上述过程中加入通道注意力，降低浅层噪点的干扰，使最终性能指标提升更加稳定。然后，通过继续引入空间注意力机制和全局对比池化从两个不同的阶段融入空间信息，使模型能够在特征提取过程中关注行人主体并抑制背景干扰。最后，在 Market-1501 和 DukeMTMC-reID

数据集上进行了实验，并通过 rank-1 和 mAP 指标验，对本章改进进行验证。

3.2 双分支行人重识别模型

近年来的行人重识别方法无论是在模型损失函数设计上，还是在特征提取粒度上，往往都不再采用单一的方法，而是使用了多种方式进行融合。这样通常能够结合多种方法的优点，获得更好的性能。在特征提取上，使用全局特征和细粒度局部特征进行融合通常能够使模型充分挖掘行人图像中的信息[72]。在损失函数设计上，使用分类损失辅助三元组损失进行训练，能够使模型在保证训练稳定的情况选获得更好的最终识别性能[1]。这几种方法在近年来的行人重识别方法研究中均被广泛应用且各具优点，本文也使用能够提取全局特征和局部特征的双分支结构作为模型的基本结构，并且使用三元组损失和分类损失进行联合对模型进行训练。接下来将对这一基本结构进行介绍。

3.2.1 模型结构

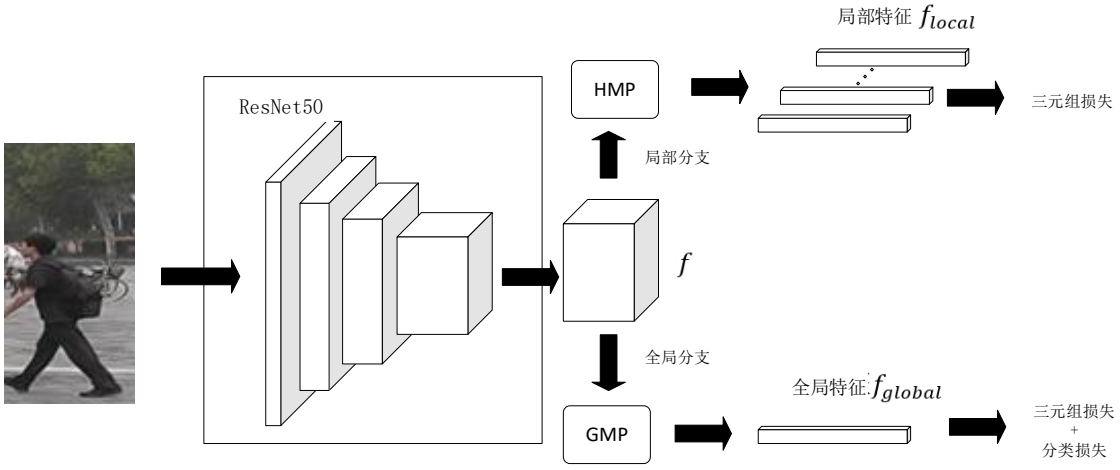


图 3.1 双分支网络模型结构

Fig 3.1 dual branch network model structure

双分支的模型结构如图 3.1 所示，图中采用的结构参考了近年来行人重识别网络的常用构建思路。本文参考 GLDFA-Net[72]的模型设计思路，模型采用 ResNet50[56]作为基线（Baseline）用于提取输入图像的基本特征描述符，并将输出的中间特征输入全局和局部分支继续进行特征提取，获取全局特征和局部特征。

在局部分支采用预定义切块的思路，将中间特征描述符$f \in R^{H\times W\times C}$，其中 H、W 和 C 分别代表特征映射的盖度、宽度和通道维度。进行垂直切分成 m 块，并进行水平最大值池化 HMP（Horizontal Maximum Pooling）成$f_i \in R^C$。最后将 m

个局部向量进行拼接得到$f_{local} \in R^{m \times C}$作为图像的局部特征。

在全局分支进行特征提取时，通常使用的方式是直接对空间维度的内的特征值进行池化降维，从而得到全局特征。行人重识别图像经过骨干网络后得到的特征描述符在 H 和 W 维度不高，在网络模型在提取全局特征时使用全局最大值池化 GMP（Global Maximum Pooling），提取空间平面内最大值最为输出。中间特征在全局分支进行池化后得到全局特征$f_{global} \in R^{C}$。

3.2.2 训练与推理

在训练阶段，行人重识别在表征学习的方法中被视为分类任务，比较常见的做法有将行人判别作为以行人 ID 为标签的分类和基于孪生网络将行人判别作为二分类两种方式。而前者在大量方法中被应用且被证明是有效且模型易收敛的。而度量学习中常用的三元组损失可以得到更具有判别行的行人特征，但是模型不易训练。根据近年行人重识别相关的研究，采用多损失联合训练可以在保证网络模型训练稳定的同时，得到更具有判别性的特征[57]。所以，本文的双分支网络采用在局部特征分支和全局分支均使用三元组损失，同时在全局使用分类损失进行联合，以加速模型的收敛增加训练平稳性。

损失函数的设计对模型训练稳定性和最终识别准确性起到至关重要的影响。在训练阶段，使用上述得到的全局和局部分支得到的特征计算相应的损失，随模型进行训练。本文中所有的模型在损失函数上均使用全局分支的分类损失、全局分支和局部分支的三元组损失进行联合训练。在模型训练时联合损失L_{total}计算式如下：

$$L_{total} = \lambda_1 L_{cls} + \lambda_2 L_{thg} + \lambda_3 L_{tl} \tag{3.1}$$

其中L_{cls}为全局分支的分类损失，L_{thg}为全局特征的三元组损失，L_{tl}为局部分支的三元组损失。λ_1、λ_2和λ_3为权重因子。L_{thg}为全局特征的困难三元组损失

在推理阶段，使用欧氏距离作为行人检索时的排序依据。其中距离包括全局距离和局部距离两部分，本文中均使用式 3-1 中两部分距离的加权和作为最终距离。

$$d_{final} = d_{global} + \alpha * d_{local} \tag{3.2}$$

其中d_{final}为最终距离，d_{global}为全局特征距离，d_{local}为局部特征距离，α是局部距离度量权重。

3.3 基于多分辨率特征融合的行人重识别方法

3.3.1 问题提出

前文提出的模型结构结合了全局和局部特征的优点，能够提取不同粒度的特征，但是将其有效地应用于实际场景仍然非常困难。这是由于深层卷积神经网络本身存在一些局限性。随着模型层次加深，在逐层进行卷积和池化操作时，如颜色轮廓这一类的低层次特征，在过程中逐渐被丢弃并用于抽取更高级的语义特征[55]。但是这一部分信息在特定情况下对行人判别是非常有效的。所以在本章节中，本文将提出了一种新的用于行人重识的多分辨率特征提取方法，用来增强模型特征的判别性。

另外根据第二章可知残差网络在特征提取过程中分为多个残差块，各残差块输出不同分辨率的特征。由于分辨率不同，中间特征无法直接相加。直接简单进行通道维度拼接的引入浅层特征可能会引入噪点信息影响模型最终的性能。而在多分辨率特征在通道维度进行融合的过程中，所以，如何不引入浅层干扰信息的前提下，融合多分辨率特征中的细粒度的特征信息是本章研究的重点。一些研究工作[59],[60]尝试将注意力引入到行人重识别方法的卷积过程中，使模型可以对关注区域进行选择，其中也包括对空间维度的选择关注的空间注意力方法。所以，本节将通过引入通道注意力解决此问题。

3.3.2 多分辨率特征融合方法

（1）多分辨率特征融合

在卷积神经网络中，输入图像沿着模型深度方向进行不断卷积和池化操作，得到分辨率逐渐降低的特征。本文将不同分辨率的卷积特征进行池化操作，使其在空间维度保持与最终层输出一致，然后在通道维度进行拼接的方式进行融合。

如图 3.2 所示，本文使用的骨干网络模型为 ResNet50，根据第二章的介绍可知骨干网络在最表层的卷积和池化之后由四个封装了多个残差单元的残差块组成。本文将四个残差块的输出逐个引出，并将由浅至深的前三个残差块的输出分别使用核大小为 8、4 和 2 的最大值池化（Max Pooling）层进行下采样，将特征描述符在空间维度上调整至于最终层的尺寸一致并在通道维度进行拼接，作为新的中间特征。

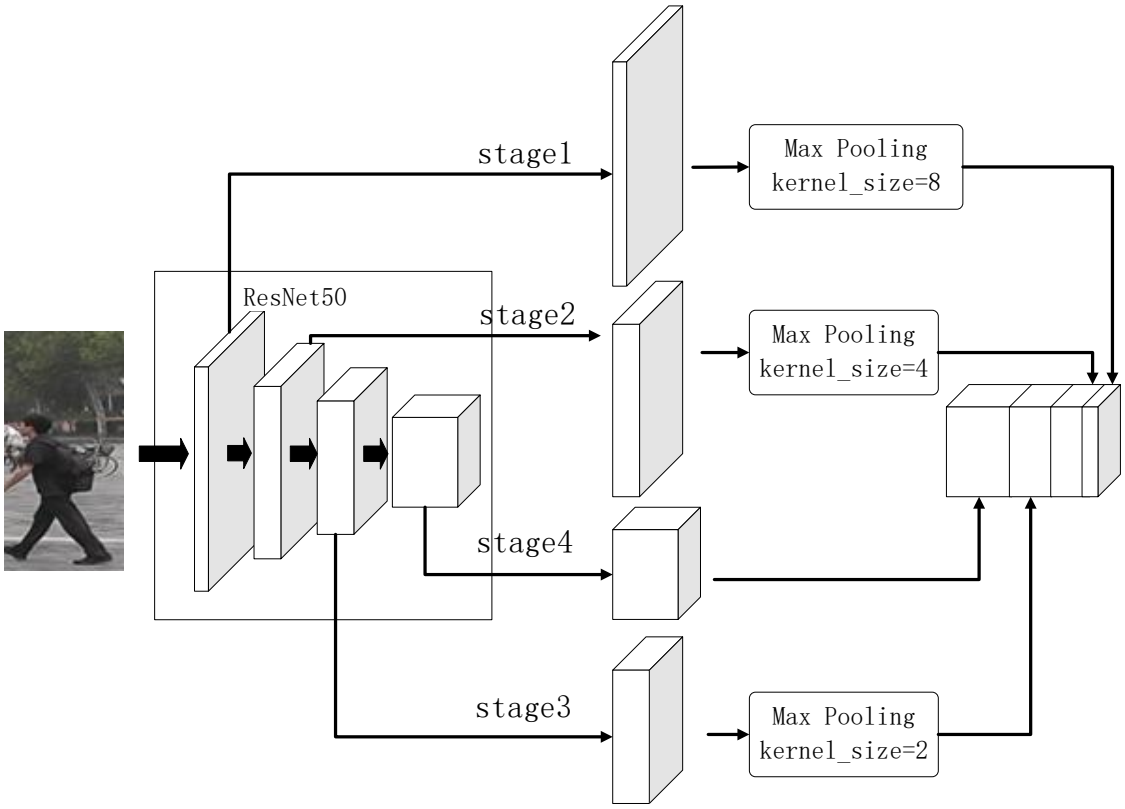


图 3.2 多分辨率特征融合网络结构图

Fig 3.2 structure of multi-resolution feature fusion network

（2）通道注意力模块

如上文中介绍可知，卷积神经网络中随着网络层次逐渐加深，卷积核的感受野逐渐增大，输出的特征语义信息更加高级。所以不同层级的残差单元输出内容也是不同的。直接简单的引入浅层特征，可能会包括浅层噪点，反而降低模型最终的性能指标。所以，需要在融合多分辨率特征中的细粒度的特征信息的同时，抑制浅层干扰信息。

如图 3.3 所示，H 和 W 分别为特征映射f的高度和宽度，C 为通道维长度。在空间注意力模块 CAM（Channel Attention Module）将中间层特征作为输入，在空间维度进行最大值和平均值池化得到$f_{cmax} \epsilon R^C$和$f_{cavg} \epsilon R^C$。然后将出输出的特征输入共享的全连接网络中，以增加结果的非线性。

其中共享全连接网络在输入和输出上保证神经元数量一致，不改变原来特征通道维度的大小。将共享网络的输出结果进行相加并使用$Sigmoid$函数进行激活，得到通道注意力值$att_{channel} \epsilon R^C$。该过程简化成公式如下：

$$att_{channel} = \sigma\left(W_2\left(\varphi\left(W_1(f_{cmax})\right)\right) + W_2\left(\varphi\left(W_1(f_{cavg})\right)\right)\right) \tag{3.3}$$

其中W_1为线性层 1，φ为 ReLU 函数，W_2为线性层 2，φ为 ReLU 函数，σ为 Sigmoid 激活函数。

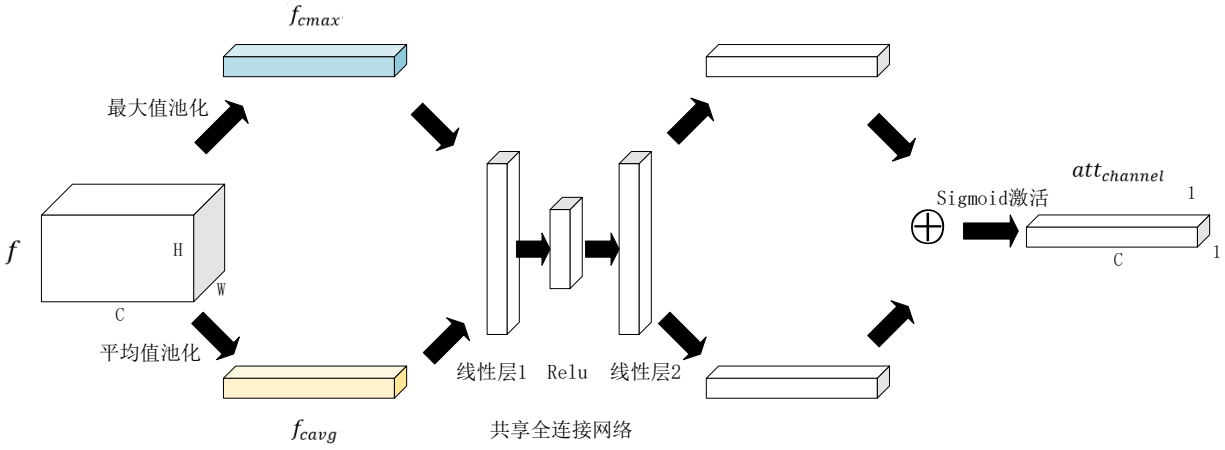


图 3.3 通道注意力模块

Fig 3.3 channel attention module

本章中多分辨率的特征采用在通道维度进行拼接的方式进行融合，所以本文引入通道注意力，让模型能够在训练过程中动态地分配各通道权重。通过通道注意力控制了不同分辨率特征在融合阶段的比重，可以小权值抑制噪点层的影响。

3.3.3 模型融合

此处将本小结提出的改进汇总成为多分辨率融合方法 MRFM（Multi Resolution Fusion Method）融合到双分支的行人重识别网络模型中。

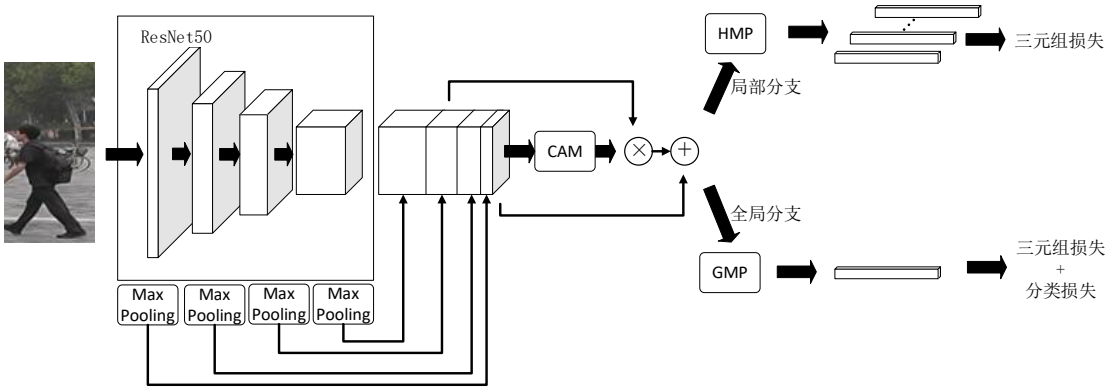


图 3.4 融合多特征的模型结构

Fig 3.4 modelstructure integrating multiple features

具体方式为如图 3.4 所示，模型在 ResNet50 的四层残差块之后将输出引出，分别通过上文所示不同尺度的最大值池化（Max Pooling）将空间维度调整至于骨干网络末端一致，并在通道维度进行拼接组成多分辨率融合的中间特征描述符。然后再将其输入通道注意力模块 CAM（Channel Attention Module）计算通道注意

力，然后与中间特征以通道乘法（Channel-wise Multiplication）方式进行融合后，再与原特征相加得到基于多分辨率融合的特征描述符，并输入下游的全局和局部分支进行后续的计算。

3.4 基于空间信息融合的行人重识别方法

3.4.1 问题提出

在行人重识别研究中，行人图像中不仅仅只有行人信息，还有背景和遮挡内容存在。在现实中，由于上游使用的目标检测算法不同，图像中的背景信息占比也随之不尽相同。若图像中背景信息过多，提取的特征中便会杂糅无用的背景信息，这在一定程度上会降低现有算法模型的性能[14]。

另外在全局分支进行全局特征提取阶段，网络会对图像的整个视觉平面进行特征提取，得到更能体现图像的整体的信息，包括明暗，颜色和整体的轮廓信息等。这一阶段，我们会对骨干网络输出的中间特征进行池化操作，将输出作为全局特征。

常见的池化方式为全局平均值池化和全局最大值池化。使用全局平均值池化得到全局特征，包含整个空间平面内的信息。使用全局最大值池化，则帮助模型聚焦于空间平面内的显著区域，从而减少因为背景信息带来的干扰。考虑到图像中行人和背景的占比可以得到如下总结，若图像中背景信息占比较大，平均值池化便会杂糅无用的背景信息；而行人占主体时，最大值池化因其取最大值的特性，势必也会丢失一些可用信息。

图 3.5 全局最大值池化和全平均值池化

Fig 3.5 global maximum pooling and global average pooling

如图 3.5 中（b）所示平均值池化通常可以够保存下整体数据的特征，其输出的结果中一般能够较好的体现出整体信息，但在行人重识别场景中会杂糅无用的背景信息过多的背景信息[73]。

而相比之下，全局最大池化往往能够更好地保存图像的纹理信息，其结果能够包含更多的边缘和轮廓信息。但是如图 3.4 中（a）所示行人信息占主体时，全局最大值池化因其取最大值的特性，势必也会丢失一些可用信息[73]。因此，本章将这两种方式进行融合，使得模型最终能够提取更具有判别性的全局特征。

3.4.2 空间信息融合方法

本节中将从中间特征提取和全局分支特征提取两个阶段，分别引入空间注意力和全局对比池化两种方法以融入空间信息，使模型能够对行人和背景进行区分，减小背景内容的干扰，从而提升模型性能。

（1）基于空间注意力进行空间信息融合

空间信息是指模型在特征提取时，可以区分图像中前景行人内容和无用背景内容的信息。行人重识别主流方法中认为平均值池得到的特征中包含更多的背景。相比之下，最大值池化更能够提取图中显著区域，随着模型训练，显著区域通常会聚焦到行人主体上。而空间信息就是利用两种池化，为模型并提供一种判断机制，使模型能够在特征提取时能够判断空间中哪个部分是有效内容。

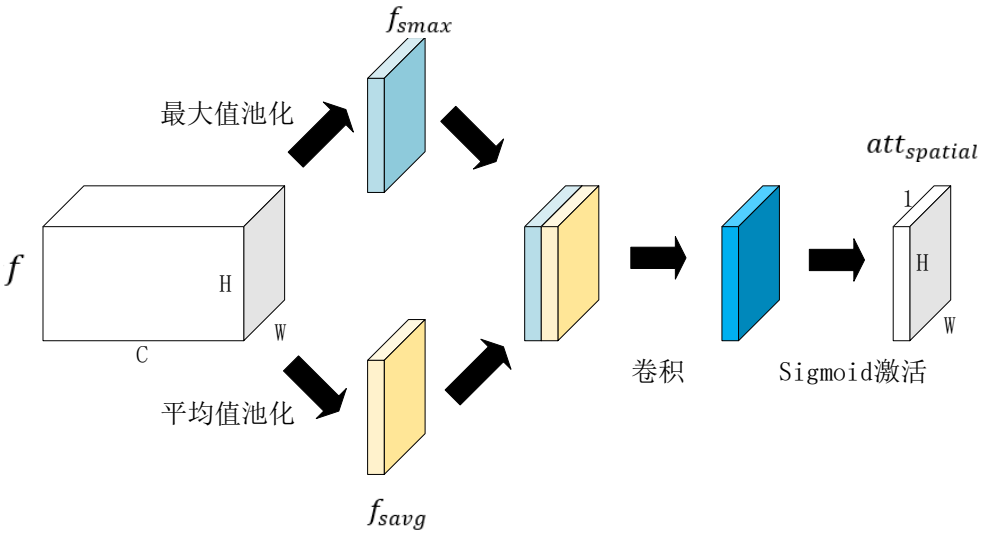


图 3.6 空间注意力模块

Fig 3.6 spatial attention module

首先在中间特征提取阶段引入空间注意力，以抑制特征中背景信息并增强模型图像中前景行人的关注。

如图 3.6 所示 H 和 W 分别为特征映射f的高度和宽度，输入的中间特征 $f \epsilon R^{H\times W\times C}$，然后在通道维度进行最大值池化和平均值池化操作，从而可以生成两个二维的特征描述符：$f_{smax} \epsilon R^{H\times W}$和$f_{savg} \epsilon R^{H\times W}$。接着我们将两个特征进行连

接并通过卷积和$Sigmoid$函数进行激活，将生成的二维矩阵$att_{spatial}$作为空间平面内的对关注区域与无效区域进行区分的注意力。该过程简化成公式如下：

$$att_{spatial} = \sigma\left(c_{2\to1}^{7\times7}\left(\left[f_{smax}; f_{savg}\right]\right)\right) \tag{3.4}$$

其中，$c_{m\to n}^{k\times k}$表示使用卷积核大小为$k*k$卷积层，并最终将输出在通道维度由m变为n，σ为 Sigmoid 激活函数，$\left[f_{smax}; f_{savg}\right]$是将$f_{smax}$和$f_{savg}$在通道维度进行拼接。

（2）基于全局对比池化进行空间信息融合

1×1卷积核，又称为网中网，因为其具有降维、升维和增加非线性的特点[62]，在一些卷积神经网络如中被广泛使用。在实现降维和升维的操作时，其实就是将通道间信息进行组合变化，可以看作是一种跨通道的信息交互。

网络最终层提取全局特征向量时，使用平均值池化和最大值池化后，其空间平面中的信息都被压缩在了最终特征向量中的每个通道值中。空间信息融合模块通过空间注意力，选择除了空间平面内的重要信息并赋予权值对前景进行强调。此处在全局分支，对前景信息进行进一步加强。

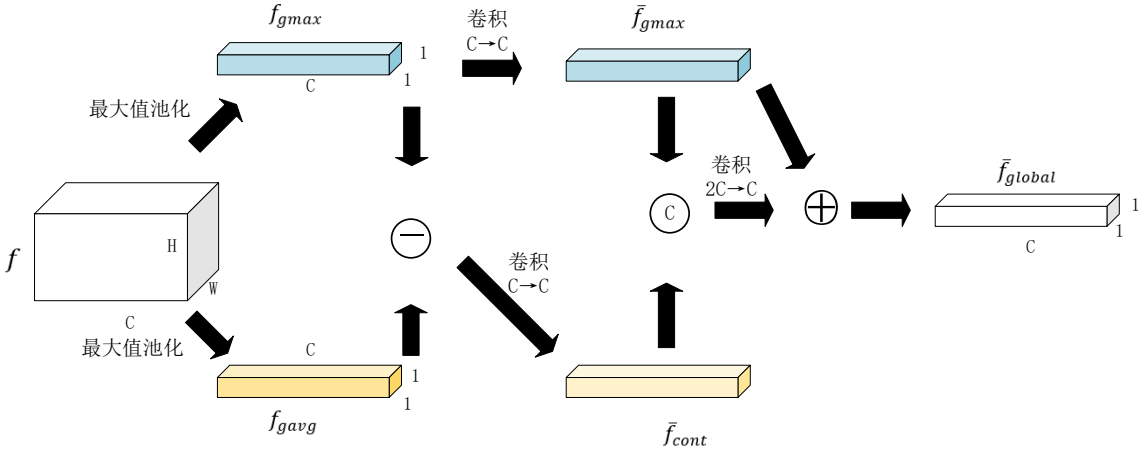


图 3.7 全局对比池化

Fig 3.7 global comparison pooling

如图 3.7 所示，将原始特征描述符$f\epsilon R^{H\times W\times C}$进行全局最大值和全局平均值池化分别得到$f_{gmax}\epsilon R^{1\times1\times C}$和$f_{gavg}\epsilon R^{1\times1\times C}$，并将其做差计算出表示平均值和最大值差异的特征向量f_{cont}。然后如式 3.5 和式 3.6 将最大值特征和差异特征分别进入不同的1×1卷积层后得到$\bar{f}_{gmax}$和$\bar{f}_{cont}$。

$$\bar{f}_{gmax} = c_{C\to C}^{1\times1}\left(f_{gmax}\right) \tag{3.5}$$

$$\bar{f}_{cont} = c_{C\to C}^{1\times1}\left(f_{gavg} - f_{gmax}\right) \tag{3.6}$$

然后将两向量进行通道维度拼接后输出通道上降维的1×1卷积层后再与$\bar{f}_{gmax}$相加得到最终的输出向量$\bar{f}_{global}$，作为最终的全局特征描述符。该过程简化

成如式 3.7 所示：

$$\bar{f}_{global} = \bar{f}_{gmax} + c_{2C\to C}^{1\times 1}\left(\left[\bar{f}_{gmax}; \bar{f}_{cont}\right]\right) \tag{3.7}$$

其中，$c_{m\to n}^{k\times k}$表示使用卷积核大小为$k \times k$卷积层，并最终将输出在通道维度由m变为n，$\left[\bar{f}_{gmax}; \bar{f}_{cont}\right]$是指将$\bar{f}_{gmax}$和$\bar{f}_{cont}$在空间维度进行拼接。

3.4.3 模型融合

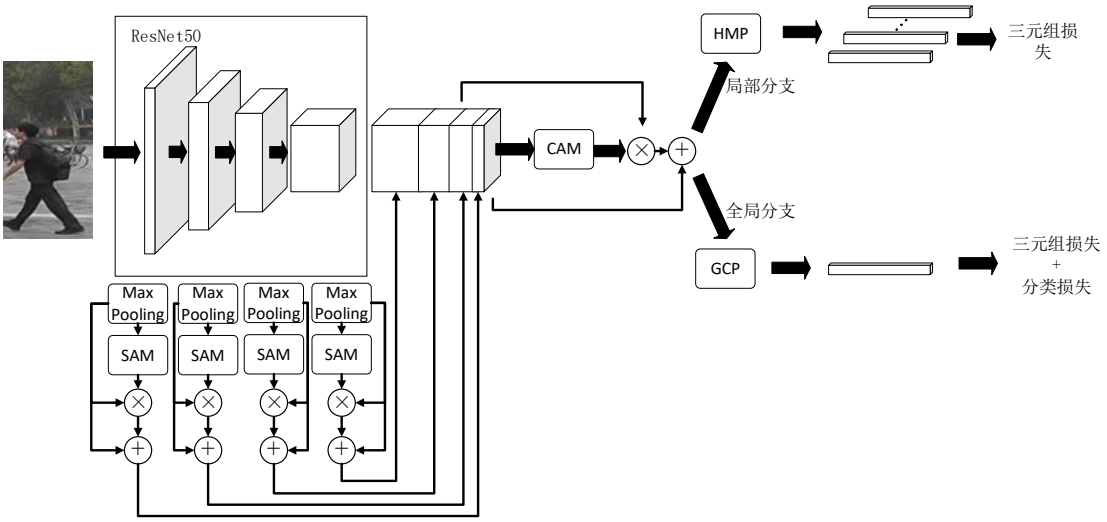


图 3.8 融合空间信息的模型结构

Fig 3.8 modelstructure integrating spatial information

本小节将提出的改进汇总成为空间信息融合方法 SIFM（Spatial Information Fusion Method）融合到上一节中优化后的双分支网络模型中。具体而言，如图 3.8 所示将在多分辨率特征进行融合之前进行空间信息融入，此处将不同分辨率的特征描述符经过最大值池化之后输入空间注意力 SAM（Spatial Attention Module）计算空间注意力。然后将其与原特征在空间维度以空间乘法（Spatial -wise Multiplication）的方式进行融合，并与原特征相加再作为后续流程输入。然后在全局分支进行全局特征提取时，使用全局对比池化 GCP（Global Contrastive Pooling）替换原来的全局最大值池化再次进行空间信息融入，并将提取的特征作为全局分支最终的特征表达用于后续损失度量。由此，得到本章最终的基于多分辨特征与空间信息融合的行人重识别模型 MFSINet（Multi-resolution Features and Spatial Information Fused Network）。

3.5 实验及分析

3.5.1 实验设置

为了验证本章提出的基于多分辨特征与空间信息融合的行人重识别模型的有效性，本章在常用的行人重识别数据集 Market-1501[63]和 DukeMTMC-reID[64]两个数据集上进行了实验，并采用 Rank-1 和平均准确率均值 mAP 进行评估，具体内容细节如下：

在本章实验中，使用 ResNet50 的主体结构，并使用 ImageNet 预训练的参数初始化，并将原始网络结构中最后一个全连接层，并将最后一阶段下采样步长设置为 1，作为特征提取的骨干网络。

为了防止过拟合，在训练过程中，进行了随机裁剪、水平翻转。图像在进入网络前都进行了缩放，统一设置为 256*128 分辨率。并在输入模型前逐通道对图像进行了标准化，使输入图像三个通道内的均值为[0.485, 0.456, 0.406]，方差为[0.229, 0.224, 0.225]，这一操作的目的是使输入图片像素值分布与 ImageNet 一致，可以使模型在训练过程中更快收敛，并提升基础网络的性能。

在训练过程中每个批次随机选取 128 个行人，每个行人选取 4 张图像，组成批次进行训练。经过骨干网络的提取出中间特征后，在局部分支将特征图划分为 8 个水平部分，每个部分经过水平池化输出特征的维度为 3840 维向量。在优化算法使用梯度下降法，使用 SGD 优化器，使用式 3.8 中设置的学习率进行训练：

$$lr(t)=\begin{cases}3.5\times10^{-5}*\dfrac{t}{20}, t\le 20\\ 3\times10^{-4}, 20<t\le 90\\ 3\times10^{-5}, 90<t\le 160\\ 3\times10^{-6}, 160<t\le 300\end{cases} \tag{3.8}$$

训练周期设置为 300。三元组的边距设置为 0.3。分类损失和三元组损失的权重由系数λ_1和λ_2进行控制初始 100 周期均设置为 1 进行同比相加，λ_3设置为 0.3。之后随训练周期呈对数趋势增大λ_2至 6.7，λ_3增大至 2.0，使其偏向于三元组损失。为了提升模型性能指标，加入了第二章提及的网络训练策略，以达到更好的重识别效果。标签平滑计算时设置ε为 0.1。在推理阶段进行距离度量时候将全局距离和局部距离系数α设置为 0.3。

本站所有算法均使用 Pytorch 深度学习框架进行实现，变成语言使用 Python3.5。所有实验基于 64 位的 Ubuntu 操作系统，硬件平台为包含英特尔 E5-2640 v4 处理器和 Nvidia GTX3090 显卡的服务器。本章实验未使用重排序算法进行指标提升。

3.5.2 评价指标

在评估行人重识别方法时，累积匹配曲线 CMC（Cumulative Match Characteristics）是最常用的评估方式，而平均准确度均值 mAP（mean Average Precision）[63]也是后来比较流行的评价指标。而现在的研究中通常会同时考虑这两个指标，以更好地评价行人重识别方法。

（1）Rank-1 指标

累积匹配曲线 CMC 曲线是模式识别系统的重要评价指标，Rank 指标关于准确率（Accuracy）的曲线，所以其本质就是观测模型 Top-k 的命中率。而在行人重识别方法的研究中，通常都简化为比较特定 Rank 下模型的识别准确率。

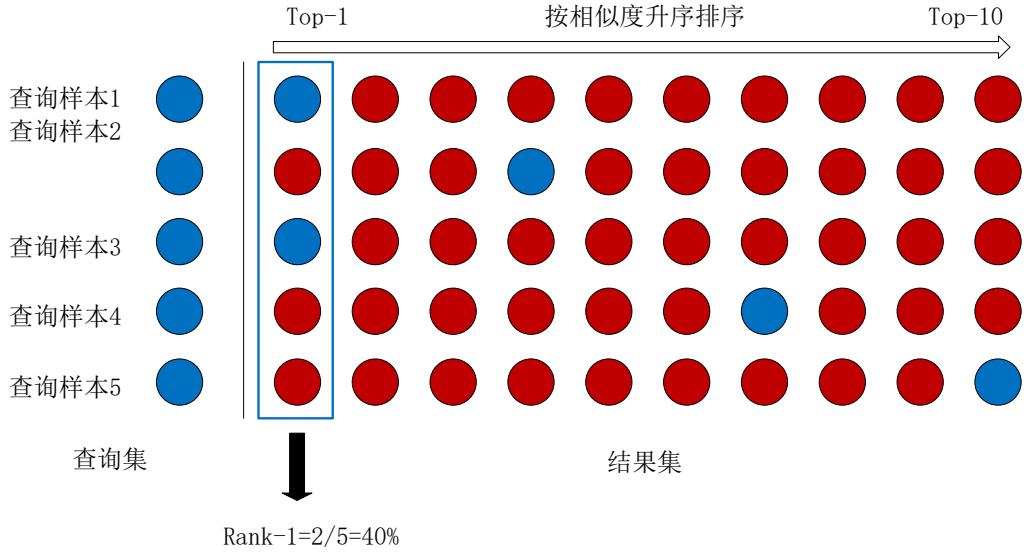


图 3.9 Rank-1 指标计算示例

Fig 3.9 example of rank-1 index calculation

算法返回的排序列表中，前 k 位为存在检索目标则成为 Rank-k 命中。如图 3.9 所示，我们假定查询集的数量为 5，在候选集中选取查询集中对应的相似度最高的 10 组候选对象，经过对比可以的到其对应比对结果。如图可知，查询集中的 1 和 3 处 Rank-1 命中，则可知算法 Ran-1 准确率为 40%。同样的可知查询集中 1、2 和 3 处 Rank-5 命中，则算法 Rank-5 准确率为 60%。同理，算法 Rank-10 的准确率能达到 100%。

（2）mAP 指标

随着行人重识别领域的发展，各模型的表现也越来越好，Rnk-k 已经能够达到比较高的准确率，仅用次指标并不能很好地区分算法间的性能。当不同的算法，都能在可接受的 k 值下得到较多的命中。此时，需要更加严格的指标，才能对方法性能进行度量。在有多个结果正确命中时，如果命中的结果在最终检索结果中排序靠前，则说明模型对正负样本的判别能力越强。

准确率是所有检索样本中检索出的正确样本的比率。结合原始标签，可以将判别结果分为四类：预测为真且实际也为真的行人图像数TP（True Positive），预测为假且实际也为假的行人图像数TN（True Negative），预测为真但实际却为假的行人图像数FP（False Positive），以及预测为假但实际却为真的行人图像FN数（False Negative）。

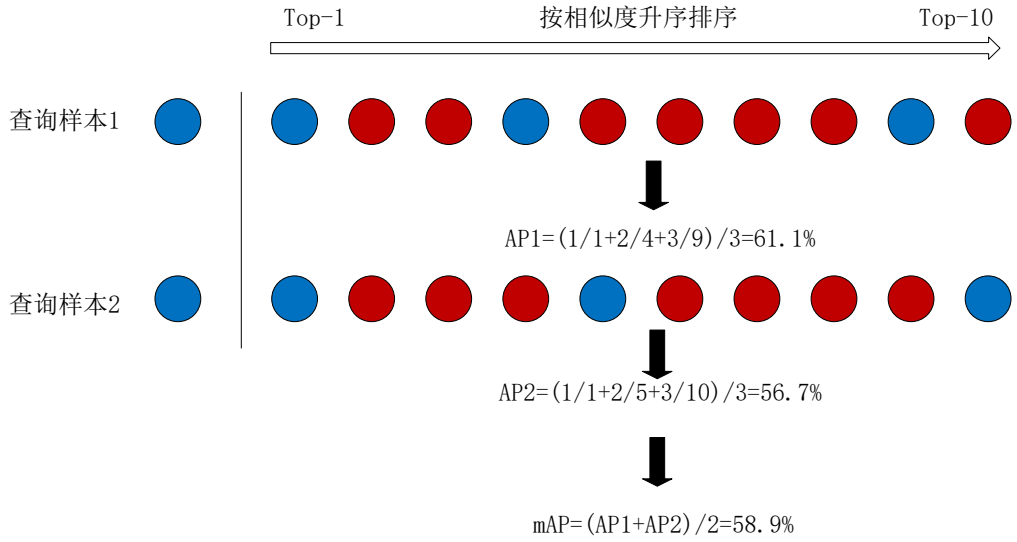


图 3.10 mAP 指标计算示例

Fig 3.10 calculation example of map index

如式 3.9 所示准确率P可以表示为查询返回的正确个数比上查询返回的总个数，即正确被识别的行人图像TP占所有实际被检测行人图像$TP + FP$的比例。

$$P = \frac{TP}{TP + FP} \tag{3.9}$$

如式 3.10 所示召回率R表示在所有正样本中有多少预测式正确的。

$$R = \frac{TP}{TP + FN} \tag{3.10}$$

那么平均的准确率AP可以表示为：

$$AP = \frac{\sum P}{C} \tag{3.11}$$

其中C指某以类别中的图像数量。在行人重识别中，会对查询集合中的所有图像计算平均准确率，则可以将其均值作模型在该数据集上的评价指标。该指标记为 mAP，计算式如 3.12：

$$mAP = \frac{\sum AP}{N} \tag{3.12}$$

其中N为样本中所有类别的数目。

如图 3.10 为 mAP 指标计算示例可知，对于查询集中的 1，在 1、4、9 处命

中，则准确率为 1/1，2/4 和 3/9 则平均准确率为 61.1%，同理 2 的平均准确率为 56.7%。而该算法相对于整个查询集，平均准确率均值为 58.9%。

本章在后续进行实验时，将同时使用 Rank-1 和 mAP 指标对融入各个方法的模型进行性能评估。

3.5.3 实验数据集

由于深度神经网络模型存在复杂的参数，需要大量的训练数据，促进了大规模行人重识别数据集的发展。所以基于深度学习的行人重识别方法研究，与探索覆盖真实场景的大规模下行人数据集密不可分。

本文选取基于图像的行人重识别数据 Market-1501[63]和 DukeMTMC-reID[64]集开展研究。

Market-1501 数据集目前使用最广泛的大规模行人重识别数据集之一，它是在清华大学一家超市前使用五个高分辨率相机和一个低分辨率相机获取的。Market-1501 使用可变形局部模型 DPM（Deformable Parts Model）[74]自动检测行人边界框。它包含 1501 个不同的行人，共有 32668 张图像，每张图像的大小为 128x64。Market-1501 标注图像多，且包含 2793+的干扰因子，更接近真实世界。

DukeMTMC-reID 数据集是在杜克大学使用八个静态高清摄像机收集的。它包含来自 702 人的 16522 张训练图像、2228 张查询图像和来自另外 702 人的 17661 张图像的候选库。同样 DukeMTMC-reID 中也存在 408 人的图像，作为干扰因子，以接近真实场景。

本章使用的数据集基本信息汇总如表 3.1 所示。

表 3.1 行人重识别数据集描述

Table 3.1 description of pedestrian re recognition data set

数据集	Market-1501	DukeMTMC-reID
行人数	1501	1812
图象数	32268	36411
相机数	6	8
检测器	DPM，手工	手工

3.5.4 实验结果及分析

为验证本章提出的基于多分辨特征与空间信息融合的行人重识别模型的有效性，分别 Market1501 和 DukeMTMC-reID 数据集上进行训练。训练过程中使用了度量和表征多损失融合联合对基准网络进行优化训练得到基线 Baseline。然

后引入第二章所提及的训练策略进行模型训练，使模型性能能够和近年来的优秀方法处于同一水平线，得到增强基线Baseline^{+}。最后以增强基线为基础，保持其他设置不变的情况下，训练本章所提出的基于多分辨特征与空间信息融合的行人重识别模型 MFSINet，并通过 Rank-1 和 mAP 对性能进行度量。

由表 3.2 为实验指标汇总，可以知道基于多分辨率特征融合和空间信息融合的方法模型对行人的判别能力有了一定程度上的提升，在 Market1501 数据集上 Rank-1 由 93.5%提升到了 94.8%，mAP 提升了也提升至 86.5%。该结果表明了在 Market1501 数据集上，采用基于本章可以更有效的提取行人特征。

表 3.2 MFSINet 性能指标

Table 3.2 Performance Indicators of MFSINet

Model	Market-1501		DukeMTMC-reID	
	rank-1	mAP	rank-1	mAP
Baseline^{+}	93.5	84.3	83.8	72.5
MFSINet	94.8	86.5	86.6	75.9

同样地，本章提出的对模型的改进在 DukeMTMC-reID 数据集上的提升相较之下更为明显，Rank-1 指标上提升至 86.6%，mAP 指标提升至 75.9%。该结果表明了在 DukeMTMC-reID 数据集上，本章所提处的方法更为有效。

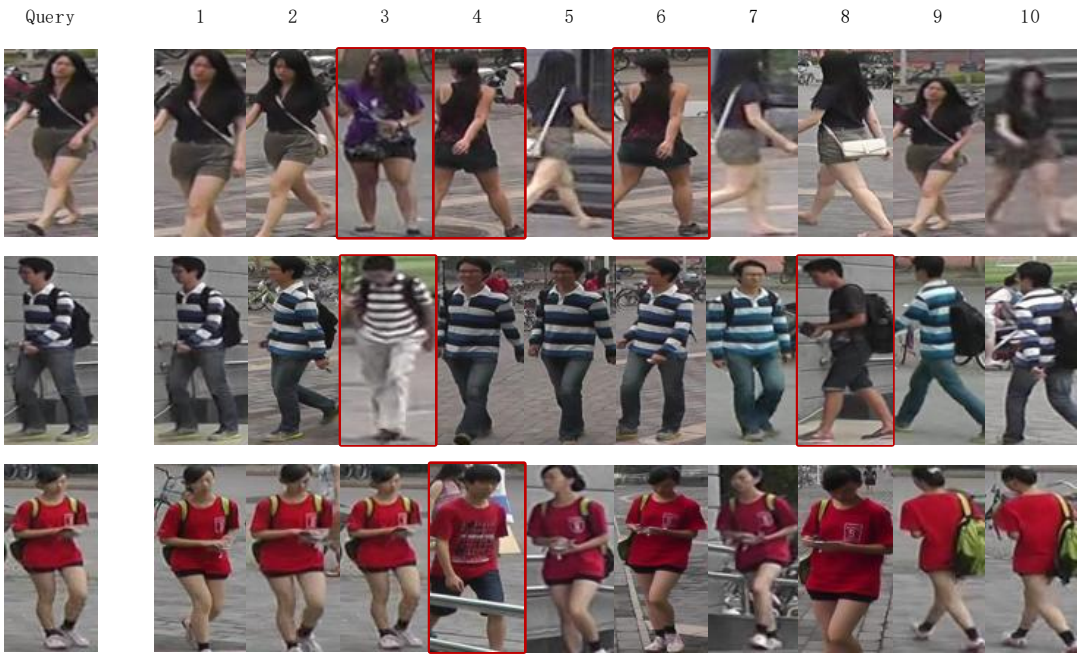


图 3.11 行人重识别查询结果

Fig 3.11 pedestrian re recognition query results

为了直观地展示本章的模型的有效性，如图 3.11 以 Market1501 为例展示了行人重识别检索效果。图中左侧为输入的查询图（Query），右侧为在候选集合

（Gallery）中查询的前十位的查询结果。图中红色标出为模型误检测结果，无框选的为正确结果。该结果表明，本章所提出的结构具有良好的检测性能。

（1）消融实验

为了验证本章改进的有效性，在 Market1501 和 DukeMTMC-reID 数据集上均进行了消融实验。具体做法为使用 ResNet50 作为骨干提出中间特征，并使用本章中的双分支行人重识别模型进行下游行人重识别，由此得到基线 Baseline。然后在 Baseline 的基础上添加第二章所述的策略（tricks）进行优化，使模型能够达到一个较高的重识别率，得到增强基线Baseline^{+}。然后通过添加本章所提出的多分辨率融合方法 MRFM 和空间信息融合方法 SIFM 并进行性能度量。由表 3.3 和表 3.4 可得本章所提出的基于多分辨率融合的方法和空间信息融合方法，均能够有效地提升模型的判别能力。

表 3.3 Market-1501 数据集上消融实验结果

Table 3.3 ablation experimental results on market-1501 dataset

Model	rank-1	mAP
Baseline	88.3	73.1
Baseline^{+}(+tricks)	93.5	84.3
+MRFM	94.3	86.0
+SIFM	94.8	86.5

由表 3.3 可得在 Market1501 数据集上，训练策略能够极大地提升模型的性能。多分辨率融合方法的加入，能够使模型提取到不同层次的更具判别性的特征，从而得到性能提升。空间信息融合方法能够在特征提取的过程中融入空间信息以降低背景的干扰，从而提升最终的识别准确率。

表 3.4 DukeMTMC-reID 数据集上消融实验结果

Table 3.4 ablation experimental results on dukemtmc Reid dataset

Model	rank-1	mAP
Baseline	80.0	65.2
Baseline^{+}	83.8	72.5
+MRFM	84.9	75.0
+SIFM	86.6	75.9

由表 3.4 可得在 DukeMTMC-reID 数据集上，通过添加训练策略同样能够提升模型的性能。在强力的增强基线，通过加入 MRFM 模块，模型提取到不同分辨率下的特征，帮助最终进行行人判别。通过 SIFM 模块降低背景的干扰的能力，模型性能得到了进一步的提升。

（2）相关方法比较

本章所提出的网络模型使用了双分支模型结构，并且同时使用了全局特征和

局部特征进行行人判别，并使用训练策略进行了性能提升并取得了比较好的性能指标，本章选择了几类方法进行性能比较。首先，以提取特征类型的维度，对基于全局特征和基于局部特征的方法分别进行了比较。然后，由于本章使用了注意力机制，又选取了基于软注意力的相关方法进行了比较。最后还与目前比较热门的基于生成对抗网络的方法进行了比较。

本章将提出的方法和改进算法与 Market-1501 数据集上的相关算法比较结果详见表 3.5。与 DukeMTMC-reID 数据集上的算法比较见表 3.6 所示。可以得出结论，本文所提出的方法是有效。

表 3.5 Market-1501 数据集实验结果对比

Table 3.5 comparison of experimental results on Market-1501 dataset

方法	Rank 1	mAP
PGFA[17]	91.2	76.8
VPM[16]	93.0	80.8
HPM[14]	94.2	82.7-
OSNet[19]	94.8	84.9
CAM[21]	94.7	84.5
HOA[65]	95.1	85.0
AANet[66]	93.9	83.4
IDE+UnityStyle[67]	93.2	89.3
MpRL[68]	85.8	67.5
CAD-Net[69]	83.7	-
Ours（MFSINet）	**94.8**	**86.5**

由表 3.5 所示，在 Market-1501 经过本章算法优化后的模型能相关方法比较中 Rank-1 指标处于同一水平。Rank-1 指标落后于基于软注意力 HOA 方法，但 mAP 指标高出该方法 1.5%。mAP 指标仅落后于基于图像风格迁移 IDE+UnityStyle 方法，但 Rank-1 指标比该方法高出 1.0%。

表 3.6 DukeMTMC-reID 数据集实验结果对比

Table 3.6 comparison of experimental results of DukeMTMC-reID dataset

方法	Rank 1	mAP
PGFA[17]	82.6	65.5
VPM[16]	83.6	72.6
OSNet[19]	88.6	73.5
CAM[21]	85.8	72.9
P2-Net[27]	86.5	73.1
IDE+UnityStyle[67]	82.1	65.2
MpRL[68]	78.8	58.6
CAD-Net[69]	75.6	-
Ours	**86.3**	**75.9**

由表 3.6 所示，在 DukeMTMC-reID 数据集上，经过本章算法优化后的模型能在相关方法比较中 Rank-1 和 mAP 指标均能处于较为领先位置。对比方法中 Rank-1 指标最高的基于托尺度特征融合的方法 OSNet，在 mAP 超过该方法 2.4%。同时 Rank-1 指标相对于基于语义信息提取的 P2-Net 落后 0.2%，但是 mAP 指标高出该方法 2.8%。综上所述，本章所提出模型具有良好的行人识别性能。

（3）多分辨率融合方法分析

为了验证多分辨率融合方法对行人重识别网络模型的性能提升，本文对各层次分辨的特征进行了多种组合，以确定最佳的融合方式。

在本章的骨干网络使用 ResNet50 的主体结构。由上文的介绍可知，本文基线保留 ResNet50 网络在表层的7×7卷积核和池化层，和后续 4 个包含 12 个卷积和的残差块作为最终的特征提取骨干网络。作为标识，将图像经过骨干中的 4 个残差块时的特征描述符分别记为f_1、f_2、f_3和f_4。

本章将通过四种特征选择方式在 Market-1501 数据集上进行实验，分别为：

（i）仅使用最终层的高级语义特征f_4；

（ii）使用最总层和较浅一层的特征特征进行融合，即使用f_4和f_3；

（iii）在上面的基础上进一步融合更浅一层的特征，即使用f_4、f_3和f_2进行拼接；

（iv）融合使用四层所有的特征，即使用f_1、f_2、f_3和f_4进行融合。

表 3.7 不同分辨率特征组合下模型性能

Table 3.7 model performance under different resolution feature combinations

f_4	f_3	f_2	f_1	Rank-1	mAP
√				93.5	84.3
√	√			94.0	85.4
√	√	√		94.3	85.8
√	√	√	√	**94.3**	**86.0**

由基于多分辨率特征融合的方法，模型对行人的判别能力有了一定程度上的提升，其中通过融合f_3特征时对模型性能提升是最大的，Rank-1 提升到了 0.5%，mAP 提升了也提升了 1.1%。通过融合f_2特征时也有一定程度上的提升，Rank-1 提升到了 0.2%，mAP 提升了 0.4%。融合最表层的f_1特征时对也是有一定的提升的，mAP 提升至 86.0%。由于表层信息较少，且在通道维度上占比最小所以提升相较于其他层的特征最不明显。

（4）空间信息融合方法可视化及分析

为了验证本章所应用的空间信息融合方法在特征提取过程中对背景信息的

抑制作用，将模型骨干部分输出的特征描述符进行了可视化，通过特征激活图展示特征关注区域。

具体做法为先将末端特征描述符进行通道维度计算平均值，再将所得的二维矩阵正则化到[0,1]之间，然后根据元素大小生成伪彩色图片。最后将生成的伪彩色图调整到原图像尺寸，并进行叠加，以展示图像关注区域。此处可视化使用 OpenCV 作为图像处理工具，在特征图转换伪色彩图像时设置 COLORMAP 为 HOT，为了获得直观的色彩差异，图像叠加时设置原始图像和伪色彩图像权值为 1.0 和-0.6。在该色彩模式中，图中红色部分表示模型关注区域，黄色部分则表示非关注区域。

图中(a)列为网络输入图像，(b)列表示基线网络提取的特征激活图，(c)列表示本章所采用的网络结构的特征激活图。由于基线网络未采用多分辨率特征，为了保证变量单一，此处统一均采用骨干网络最后一层输出的特征描述符进行可视化展示。

(a)　　(b)　　(c)

图 3.12 特征激活图

Fig 3.12 feature activation diagram

由图 3.12 所示，(b)列表示基线网络提取的特征激活图中，模型在全局特征提取时能够对图像中有用的判别区域进行激活，但是所有通道的激活信息进行平均之后可以明显看出其中包含了较大范围的背景区域。由于最终层特征在空间维度上分辨率较小仅有 16*8，所以关注区域可视化之后粒度较粗。但是(c)列表示本章所采用的网络结构的特征激活图中，已经可以明显看出图中激活区域更加集

中于行人区域，其中背景部分占比有所降低。综上所述可知本章所提出的空间信息融合方法，在一定程度上可以使模型聚焦图像前景，关注行人身上具有判别性的部位，减少背景信息的干扰。

3.6 本章小结

本章基于多分辨率特征融合的行人重识别方法。首先，针对现有的网络结构，结合卷积神经网络的特点，描述了模型在终层对底层特征信息确实的问题，并提出了基于多分辨率特征融合的思路；并介绍了行人重识别中常见的样本背景复杂的问题，并结合当前的网络结构，提出引入了空间信息融合模块以解决背景信息干扰的问题。最后，对提出的改进进行了实现，并在开源数据集上进行实验，验证了改进的结构的有效性。

第 4 章 基于局部特征融合的切块对齐行人重识别模型

4.1 研究动机

行人重识别方法中局部特征能捕捉比单纯的全局特征更细粒度的判别信息，从而可以获得更好的判别效果[70]。但是，基于局部特征的行人重识别方法对于图像中行人对齐要求比较高。而在实际生产应用中，行人图片由于分辨率各异、行人姿态角度变化、相机焦段拍摄角度不一以及受各种前景遮挡等问题，均会导致行人在空间上拓扑结构不对齐，从而影响行人重识别模型的性能指标。

基于预定义大小图像切块是一种低成本且高效的局部特征提取方法[15]。行人重识别方法的输入是经过上游目标检测得到的行人图像。人体在垂直方向上的拓扑结构，大致可分为头部、上身、躯干、下肢等[1]。所以基于切块的局部特征提取方法，通常以垂直切割的方式，将图像分割为沿垂直方向排列的多个水平图像块。但是，这种简单均匀切分并不总能在图像块中得到均匀一致的人体结构。Sun [15]等人通过在 PCB 模型中进一步引入 RPP 对硬切分进行微调，以降低不对齐问题的影响。但是，该方法只关注了局部特征，并没有充分利用全局信息。Luo[57]提出 Alignedreid 使用图像各个切块之间的局部距离，通过最短路径在局部距离矩阵中搜索最佳的距离度量链路。但是这种方式只考虑了最相似的切块之间的距离，舍弃相邻的其他部分，相当于一种硬对齐的。

为了解决不对齐问题对模型性能的干扰，本章将在第三章的模型基础上进行改进，提出基于局部特征融合的切块对齐行人重识别模型。首先，针对不对齐问题对行人重识别数据集中的图像进行了细粒度分类，并以此对模型误判过程进行了详细的分析。然后在局部分支根据切块间相似度，对原始样本在局部特征层面进行本地对齐计算，以减少图像不对齐引入的距离度量误差。另外，通过双分支信息融合的困难样本挖掘方法，在不引入过多计算量的情况下，使局部特征分支在模型训练过程中获得更多造成不对齐问题的样本。并在 Market-1501 和 DukeMTMC-reID 数据集上进行了实验，并通过 rank-1 和 mAP 指标验，通过并详细的可视化工作对算法原理性进行原析。

4.2 基于预定义切块的局部特征提取方法

本章方法针对解决局部分支在特征提取阶段由于行人垂直方向拓扑结构不对齐，引起的训练阶段不对齐样本对局部损失过大和推理阶段局部距离过大的问题。本文中采用的基于切块的局部特征，本节首先详细介绍局部特征提取过程。

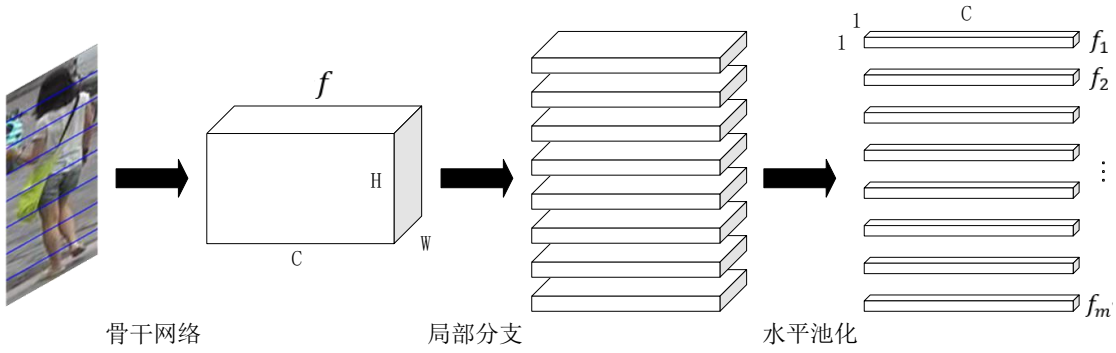


图 4.1 局部特征提取方法

Fig 4.1 local feature extraction method

如图 4.1 图像经过骨干网络进行特征提取，得到包含图像信息的三维特征表达式$f \epsilon R^{H \times W \times C}$。随着图像原始的张量在网络中进行卷积和池化操作，其空间上维度的长度减少，通道维长度增加。根据卷积神经网络的计算特性，同一通道的特征值是由同一卷积核提取而来，且与原图像相对位置区域的图像内容相对应。所以可以在特征层面，通过对空间相关维度进行分割，得到原图形对应区块的局部特征图。

根据行人重识别图像的特点，通过垂直切割得到的依竖直方向均匀排列的图像块，更符合人体拓扑结构的特点，所以更适合用来提取局部特征。所以本文中也局部特征提取均采用此方法，将行人进行竖直方向的均匀m等分。切分之后，将原始特征图在W维度进行水平最大值池化，假设C维长度为n，便可以得到m个长度n的特征向量。

4.3 局部特征融合的切块对齐方法

4.3.1 问题提出

本文在进行局部特征提取时，采用上文提及的在垂直方向上基于预定义切块的方式进行行人局部特征提取。但是在这种方式在局部特征提取并进行局部距离度量时，需要保证行人在空间中的拓扑结构有较强的对应关系。但是在实际的场景中，行人图像来自上游的行人检测系统。如本文使用的 Market-1501 和 DukeMTMC-reID 数据集，前者使用 DPM 和手工的方式进行行人框定，后者使

用手工的方式进行框定。其框定出的行人图像结果并不能保证行人在图像中始终占据相同的大小和比例。所以会出现在垂直方向上行人图像不对齐的现象。

(a)　　(b)

图 4.2 Market1501 图像样本示例

Fig 4.2 example of market1501 image sample

如图 4-2 所示，（a）展示了垂直方向上对齐的但是不属于同一行人的图像；（b）展示了属于同一目标，但是在空间结构上不对齐的一组行人图像。

随着姿态估计模型的发展，将姿态模型引入行人重识别中，获取行人部位局部特征，以解决行人局部不对齐问题[25]，成为当前的一个热门研究方向。但是，此类研究需要引入额外的模型，模型结构复杂。所以如何从这种样本间进行低成本的对齐，减小局部不对齐的影响是本文研究的重点。

此外，本文在全局和局部分支均使用了三元组损失进行模型优化。在模型训练时，采用基于批次的方式进行数据输入。基于批次的困难样本采样，能够使模型在训练时动态地在本地进行困难样本筛选，在引入较小的计算量的同时使模型能拥有更好的性能。但是双分支结构下独立进行困难样本筛选，将会在训练阶段因为筛选结果不一致导致模型训练不稳定。而且，会引入较大的计算量。所以还需要在训练阶段双通道距离度量流程中，对局部分支采样进行优化，使局部分支在训练阶段获得更多会引起不对齐误差的样本，强化模型对空间不对齐行人图像的判别能力。

4.3.2 问题分析

在行人重识别模型进行训练时局部特征分支使用三元组损失对模型进行优化，样本可分为正样本和负样本。由于局部特征分支使用基于预定义切块的方式进行局部特征提取，图像中人体在垂直方向上的拓扑结构是否对齐对于计算三元组损失时距离度量影响巨大。

本章结合行人样本的特点，可将正负样本进行进一步的更细地分类。如图 4.3

分别为对齐的正样本 ap（Aligned Positive Samples），不对齐的正样本 up（Unaligned Positive Samples），对齐的负样本 an（Aligned Negative Samples）和不对齐的负样本 un（Unaligned Negative Samples）。

图 4.3 行人样本对细粒度分类

Fig 4.3 fine grained classification of pedestrian sample pairs

其中对齐的正样本和对齐的负样本在局部距离度量时不会因人体的拓扑结构引入额外误差。如图 4.4 不对齐的负样本在特征空间中，理论上距离锚点最远，对于排序得到的行人重识别结果影响是最小的。而处于中间位置的对齐正样本，由于受到空间不对齐因素的影响，在特征空间中相对于对齐正样本更加远离锚点，容易在排序过程中与对齐的负样本进行混淆，从而造成模型误判。

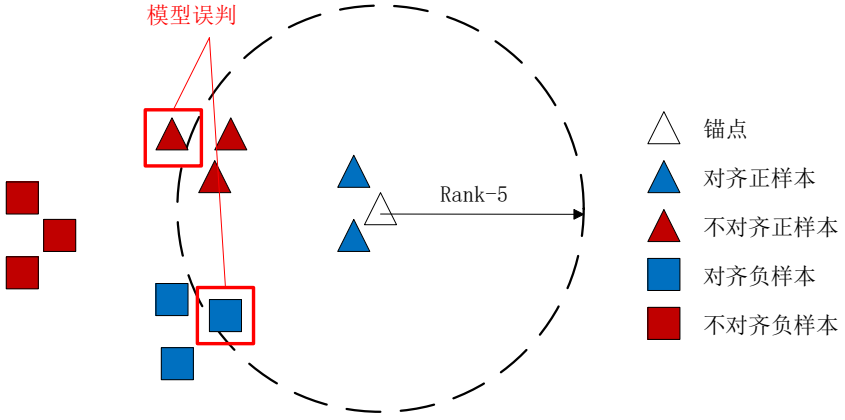


图 4.4 模型误判原理分析

Fig 4.4 analysis of model misjudgment principle

如图 4.5，解决此问题的关键点在于保证其他样本间距离一定的同时，缩进不对齐正样本之间的距离。如何凸显不对齐的正样本和对齐的样本之间的差异，是本章算法的主要思路。

本章后续将提出基于局部特征融合的切块对齐方法，削弱因空间不对齐误差导致的不对齐正样本和对齐负样本之间的混淆。并会在不同数据集上实验并进行可视化操作，以验证此思路的有效性及进行进一步的原理剖析。

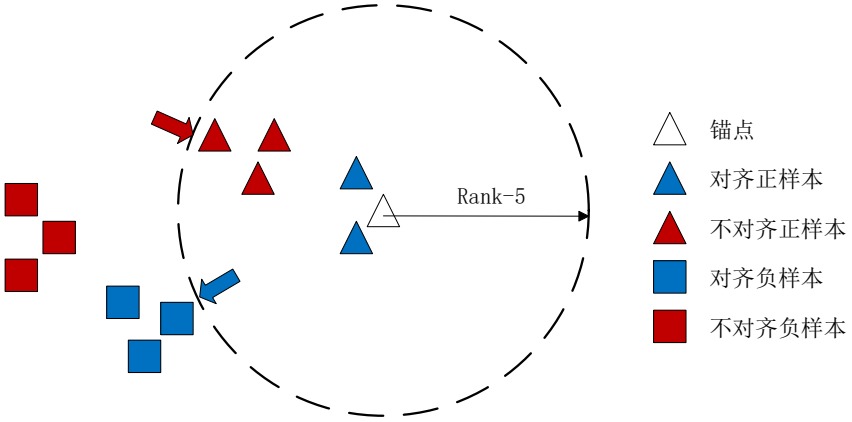


图 4.5 优化思路

Fig 4.5 optimization ideas

4.3.3 切块对齐方法

（1）原始距离度量

正如前文演示的不对齐正样本对 up，因为人体拓扑结构在垂直方向上不对齐，会得到过大的特征距离，从而可能被模型视为负样本而引入误差。本节将图 4.3 中的样本对 ap 中的两张行人图像经图 4.1 所示孪生的骨干网络的进行中间特征提取，然后经过切块和水平池化，得到m个n维局部特征，组成$m \times n$的特征矩阵记分别为Q和G表示如下：

$$Q = \begin{bmatrix} q_{11} & q_{12} \cdots & q_{1n} \\ q_{21} & q_{22} \cdots & q_{2n} \\ \vdots & \vdots & \vdots \\ q_{m1} & q_{m1} \cdots & q_{mn} \end{bmatrix} = \begin{bmatrix} \overrightarrow{q_1} \\ \overrightarrow{q_2} \\ \vdots \\ \overrightarrow{q_m} \end{bmatrix} \tag{4.1}$$

$$G = \begin{bmatrix} g_{11} & g_{12} \cdots & g_{1n} \\ g_{21} & g_{22} \cdots & g_{2n} \\ \vdots & \vdots & \vdots \\ g_{m1} & g_{m1} \cdots & g_{mn} \end{bmatrix} = \begin{bmatrix} \overrightarrow{g_1} \\ \overrightarrow{g_2} \\ \vdots \\ \overrightarrow{g_m} \end{bmatrix} \tag{4.2}$$

此处我们需要定义水平块之间的距离度量函数，此类函数定义只要能够在特征空间内能够区分出样本的相似度即可。这类函数有很多，也可以根据需要定义，但是在深度学习端对端的训练中，通常我们会使用欧氏距离、马氏距离、曼哈顿距离或者是余弦相似度等尽可能连续可导的函数。此处选择与全局分支一致的欧式距离来度量，特征空间内距离远则表示相似度低，反之则高。

定义沿垂直方向上图像 1 的第i块和图像 2 的第j块的距离为：

$$d_{ij} = \|\overrightarrow{q_i} - \overrightarrow{g_j}\| \tag{4.3}$$

则，样本间的距离为：

$$d_{Q,G} = \sum_{i=1}^{m} d_{ii} \tag{4.4}$$

图 4.3 中正样本对 up 所示，因为人体拓扑结构在垂直方向上不对齐，会得到过大的特征距离，从而可能被模型视为负样本而引入误差。

（2）局部切块特征对齐

卷积神经网络，会在卷积计算是从低到高不同视角下提取图像的特征。所以在切块内内容越相同，则切块间的距离也越近。如图 4.6 可以通过这一特性，对图像切块进行重新排列，来降低这一误差的影响。本节中介绍计算对齐矩阵和切块对齐的具体流程。

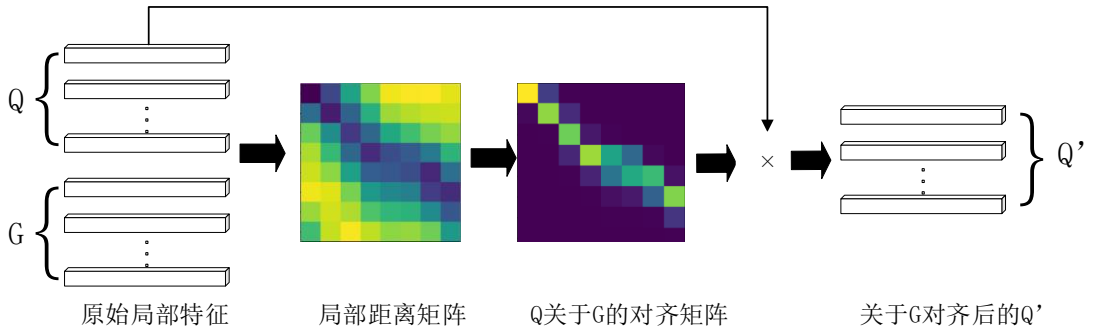


图 4.6 局部特征融合模块

Fig 4.6 local feature fusion module

首先可以通过式 4.3 计算两图像之间各个切块的距离，并组合得到局部距离矩阵为：

$$D_{Q,G} = \begin{bmatrix} d_{11} & d_{12} \cdots & d_{1m} \\ d_{21} & d_{22} \cdots & d_{2m} \\ \vdots & \vdots & \vdots \\ d_{m1} & d_{m1} \cdots & d_{mm} \end{bmatrix} \tag{4.5}$$

局部距离矩阵，反应了两图像各切块之间的距离关系，我们需要通过距离关系提取出需要的对齐信息，然后在特征层面进行对齐处理。此处实例中，我们选择G作为参照，将Q进行垂直向的特征对齐。通过距离关系可以计算出各图像块之间的相似度作为注意力值，得到两图像之间的注意力矩阵。将$[d_{11}, d_{21}, \ldots, d_{m1}]^T$记为$\overrightarrow{d_{g1}}$，表示$G$的第一各切块到$Q$各切块的距离。由于距离越小则相似的程度越高，则切块间的注意力定义如下：

$$\overrightarrow{att_{g1}} = Softmax\left(-\lambda_{att} * \frac{\overrightarrow{d_{g1}} - Mean\left(\overrightarrow{d_{g1}}\right)}{Mean\left(\overrightarrow{d_{g1}}\right)}\right) = [a_{11}, a_{21}, \ldots, a_{m1}]^T \tag{4.6}$$

其中：

$$Softmax(xi) = \frac{e^{x_i}}{\sum_{c=1}^{C} e^{x_c}} \tag{4.7}$$

令$p(q_i|g_1) = a_{i1}$称为以g_1为参照，重组q_1时选取q_i的概率，也称之为g_1对于Q中各切块的注意力。则可以根据g_1的注意力向量，在特征层面对q_1进行重组对齐，得到对齐后的特征向量为:

$$\overrightarrow{q_1}' = \sum_{i=1}^{m} \overrightarrow{q_i} * p(qi|g1) = \overrightarrow{att_{g1}} \times Q \tag{4.8}$$

由式 4.6 可知，若对于整个Q进行重组，则先计算以G为参照Q的重组注意力矩阵，计算式如下：

$$Att_{Q,G} = Softmax\left(-\lambda_{att} * \frac{D_{Q,G} - Mean\left(D_{Q,G}\right)}{Mean\left(DD_{Q,G}\right)}\right) = \begin{bmatrix} a_{11} & a_{12} \cdots & a_{1m} \\ a_{21} & a_{22} \cdots & a_{2m} \\ \vdots & \vdots & \vdots \\ a_{m1} & a_{m2} \cdots & a_{mm} \end{bmatrix} \tag{4.9}$$

其中$Softmax$在计算时需要设置计算维度为0，$Mean$函数表示矩阵中数值均值。

获得注意力矩阵之后，以G为参照对Q进行融合对齐，就是以G中每个切块对Q的切块进行加权相加，将结果作为对齐后的特征。

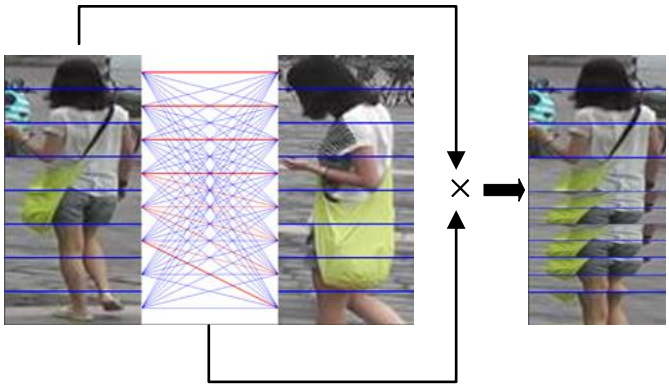


图 4.7 局部特征融合过程

Fig 4.7 local feature fusion process

如图 4.7 为原始图像进行对齐的思路，图中切块之间连线为对齐注意力矩阵。在图像切块之间设定 0.25 为阈值，低于此阈值设为蓝色，高于此阈值设为红色，并设置图像连线宽度与注意力值成正比进行切块间对齐可视化。则红色线条连接相似切块，蓝色连接非相似区域。通过$Att_{Q,G}$进行融合后，得到最右图像。图中

为思路演示，实际计算时在局部特征层面进行，具体计算过程如下：

$$Q' = Att_{Q,G}^{T} Q = \begin{bmatrix} q'_{11} & q'_{12} \cdots & q'_{1n} \\ q'_{21} & q'_{22} \cdots & q'_{2n} \\ \vdots & \vdots & \vdots \\ q'_{m1} & q'_{m2} \cdots & q'_{mn} \end{bmatrix} = \begin{bmatrix} \overrightarrow{q'_1} \\ \overrightarrow{q'_2} \\ \vdots \\ \overrightarrow{q'_m} \end{bmatrix} \tag{4.10}$$

（3）对齐后的距离度量

根据Att对Q进行重组对齐，得到重组后的特征矩阵Q'后。重组后样本对切块间的距离为：

$$d'_{ij} = \left\| \overrightarrow{q'_i} - \overrightarrow{g_j} \right\| \tag{4.11}$$

则重组后得到的的Q'与G的距离为：

$$d'_{Q,D} = \sum_{i=1}^{m} d'_{ii} \tag{4.12}$$

作为最终样本对的局部分支距离。

（a）原始样本对　　（b）对齐样本对

图 4.8 局部特征融合前后样本对

Fig 4.8 sample pairs before and after local feature fusion

如图 4.8 所示，局部特征融合前两样本之间由于空间上不对齐，在进行距离度量时会引入额外的误差。局部特征在垂直空间上进行对齐后，一定程度上可以缩小了这一误差对局部距离度量的影响。

4.3.4 分支信息融合的双分支困难样本挖掘方法

本文的方法中，在局部分支和全局分支都是用了度量学习的方法对模型进行训练。度量学习的过程本质其实就是一个映射$f(x): R^F \to R^D$，将原始样本从原空间映射到特征空间。然后再使用函数$Dist(x,y): R^D \times R^D \to R$对距离进行度量。然后根据定义的损失函数，以最小化损失为目标，寻找最优模型参数。本节中将对这一损失计算过程进行优化，使模型在训练过程中接触更多会引起不对齐问题的样本，从而增强对这一类样本的判别能力。

（1）三元组损失

根据直观的理解，模型最终确定的特征空间需要满足正样本分布呈簇，负样本的簇之间远离至一个可区分的距离。则可以定义损失：

$$L_c = y d_{I_1,I_2}^2 + (1-y)(\alpha - d_{I_1,I_2})_+^2 \tag{4.13}$$

其中：

$$z(x)_+ = \max(z, 0) \tag{4.14}$$

对比损失中样本对的选择时正负样本对的比例，会直接影响最终的模型训练效果。为了可以同时兼顾正负样本损失，目前行人重识别领域使用最广泛的度量损失为三元组损失。三元组损失中每次输入模型的都是随机采样的三张样本，其中包括锚点样本a，来自同一行人的图像p和来自不同行人的图像n，构成正样本对$\{a,p\}$和负样本对$\{a,n\}$。三个样本作为一组，经过相同的模型映射f，再经过$Dist$函数进行距离度量后，就可以得到正样本对距离$d_{a,p}$和负样本距离$d_{a,n}$,然后定义损失函数为：

$$L_t = (\alpha + d_{a,p} - d_{a,n})_+ \tag{4.15}$$

其中$d_{a,p}$表示样本a与随机正样本之间的距离，$d_{a,n}$表示样本a与随机负样本之间的距离。

这样可以保证每次通过损失函数进行反向传播，更新参数时都可以兼顾正负样本，从而模型训练更稳定。

但是这些损失在进行正负样本选取的时候，都是基于样本标签进行的随机采样。我们定义相似的正样本为简单正样本，极为不同为简单负样本，统称为简单样本。而使用随机采样的方法，使得模型输入的样本中包含大量的简单样本，最终导致在训练阶段能够达到很好的效果，但不具有泛化性。

（2）基于批次的困难三元组采样法

由于普通三元组损失的缺点在于随机从训练集中挑选三张图片，那么可能挑选出来的可能是很简单的样本组合，本文在全局分支已经引入基于批次样本的困难样本采样方法，本小节中对具体的计算流程进行描述，并在下以小节对局部分支进行优化。

在实际训练时通常采用批次训练而非仅仅使用一个三元组。假设我们选取N个目标，从属于每个目标的图像中选取M张图像，共$N \times M$张图像集合作为一个批次输入到我们的模型中。其中对于每个样本a而言，存在$M-1$张图像p可与其组成正样本对的集合P，$(N-1) \times M$张图像n可与其组成负样本对组成集合N。

可以定义与a最不相似的样本p为基于批次的困难正样本，与a最相似的负样本为基于批次的困难负样本，统称为基于批次的困难样本。

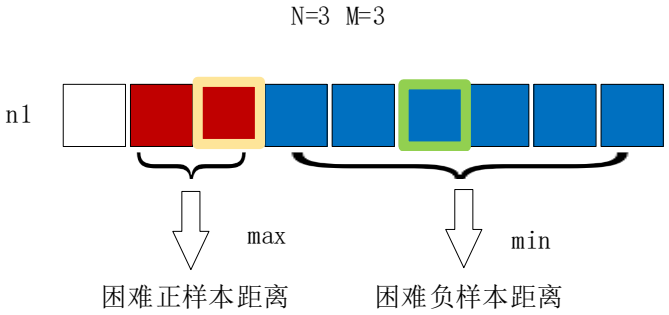


图 4.9 困难样本挖掘过程

Fig 4.9 difficult sample mining process

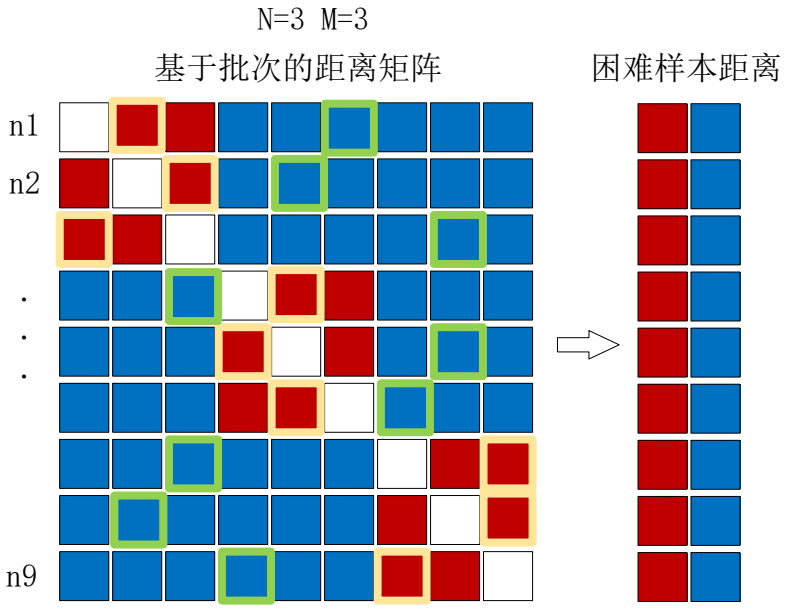


图 4.10 批次困难样本挖掘过程

Fig 4.10 mining process of batch difficult samples

如图 4.9 所示为N为 3，M为 3 时单个样本困难样本挖掘过程。如图 4.10 为整个批次进行样本采样的结果。图中所示白色部分处对应样本与自身的距离，红色部分为与对应位置正样本之间的距离，蓝色部分为与对应位置处负样本之间的距离。黄色框选处为困难正样本挖掘机之后的索引，绿色部分为困难负样本索引。

此前已经介绍，本文使用样本的在特征空间的距离作为差异性的度量，而正样本之间差异较小，而负样本之间差异性较大。因此基于此，选取距离最大的正样本作为困难正样本，而距离最小的负样本作为困难负样本。则基于困难采样的三元组损失为：

$$L_{ht} = \frac{1}{N * M} \sum_{a \in B} (\alpha + \max(d_{a,p}) - \min(d_{a,n}))_+ \tag{4.16}$$

其中$d_{a,p}$表示样本 a 的困难正样本距离，$d_{a,n}$表示困难负样本距离。

（3）双分支困难样本挖掘信息融合

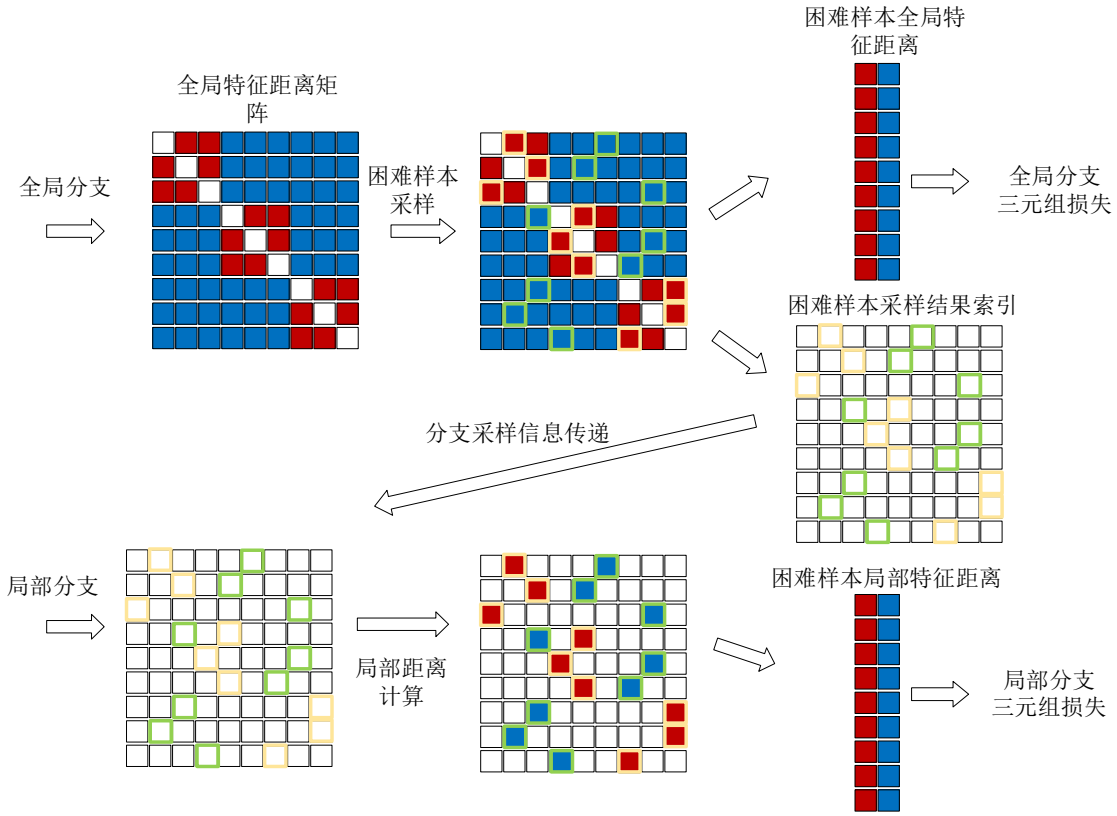


图 4.11 双分支困难样本挖掘

Fig 4.11 double branch difficult sample mining

全局分支和局部分支此前都使用了基于三元组损失的度量学习，但是若直接将两分支的三元组损失进行直接替换，在实际实现过程中仍有许多问题。

（a）基于以上的困难样本挖掘方法，需要计算每一对样本之间的距离，然后计算从中筛选困难样本。而在局部分支对样本进行距离度量之前需要进行空间对齐，其在计算上是串行的两个过程，会在此采样阶段引入过大的计算量。

（b）使全局分支分别进行困难样本挖掘时，由于采样结果可能有所不同，模型在训练阶段不稳定。并且全局分支使用三元组损失和分类损失共同优化模型，且分类损失在早期就能够帮助模型快速。所以模型训练早期会偏向于优化全局分支，局部特征此时单独进行困难样本挖掘效率低下。

基于以上原因，本文中尝试如图 4.11 所示在全局分支进行完整的困难样本挖掘，然后将全局分支的困难样本索引传递到局部分支。局部分支仅针对索引选择样本进行局部距离度量，等同于在局部分支实现低成本的困难样本挖掘。

如图 4.3 中所示的样本细粒度分类中可知，不对齐的正样本与对齐正样本相比，不对齐正样本与锚点样本之间存在更明显的差异。而对齐的负样本与不对齐

的负样本相比，对齐负样本与锚点样本更加相似。如式 4.16 所示，全局分支困难样本挖掘过程中，会倾向于筛选具有差异的正样本作为困难正样本，更加相似的负样本作为困难负样本。在训练过程中，会筛选出更多的不对齐样本，用于模型训练。经过上文的分析，不对齐问题对算法判别性能的主要影响来自于不对齐的正样本和对齐的负样本，而对齐算法也主要针对这一类样本进行区分。所以本节中的方法，在局部分支的困难样本挖掘的方向，以上文算法的优化目标是一致的。

为了在不引入额外计算量的情况下对局部分支进行困难样本挖掘，实际计算方法如图 4.11 所示。首先在全局分支计算出基于批次的样本全局特征距离矩阵，然后利用上述基于批次的困难样本挖掘方法对全局分支进行困难样本采样。全局分支的采样结果又两个输出产物，分别为基于批次的困难正负样本距离向量和困难样本采样结果索引。其中困难正负样本距离向量用于计算整个批次的困难三元组损失，采样索引则传递给局部分支。局部分支则根据采样索引直接在批次中采样对应样本的局部特征并进行距离度量，并作为后续局部分支的困难三元组损失的输入。这样即保证了双分支的采样结果一致，又减少了局部分支的计算量。

加入双分支困难样本挖掘后，模型训练时最终的损失L_{total}计算表达式改写为：

$$L_{total} = \lambda_1 L'_{cls} + \lambda_2 L_{thg} + \lambda_3 L'_{thl} \tag{4.17}$$

其中L'_{cls}为全局分支标签平滑的分类损失，L_{thg}为全局特征的困难三元组损失，L'_{thl}为对齐的局部特征融合了全局困难样本采样信息后的困难三元组损失，λ_1、λ_2和λ_3为权重因子。

4.3.5 模型融合

如图 4.12 所示本章将双分支困难样本挖掘方法 DBHM（Dual Branch Hard Mining）和局部特征融合的对齐方法融入到前文的模型结构中。在全局分支三元组损失度量之前进行全局困难样本挖掘 GHM（Global Hard Mining）。然后将采样索引传递到局部分支进行局部困难样本挖掘 LHM（Local Hard Mining），局部分支将根据困难样本输入局部特征融合模块 LFFM（Local Feature Fusion Module）进行局部特征对齐处理，最后进行局部通道的三元组损失计算。融合了本章方法后得到切块对齐网络 SANet（Slice Alignment Network）作为最终的网络结构。

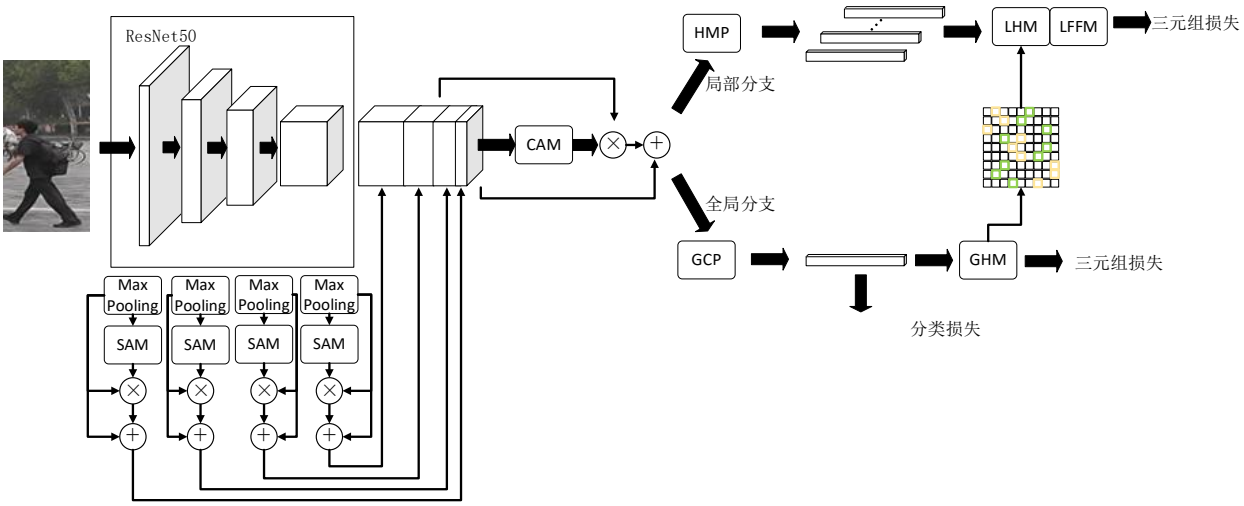


图 4.12 基于局部特征融合的模型结构

Fig 4.12 model structure with local feature fusion

4.4 实验及分析

4.4.1 实验设置

为了验证本章提出的模型改进方法有效性，本章同样在在 Market-1501 和 DukeMTMC-reID 两个数据集上进行了实验，采用与第三章一致的 Rank-1 和 mAP 进行评估。

本章的实验基于第三章 MFSINet，并添加基于特征融合的切块对齐方法。批次中每个批次的训练随机选取 128 个行人，每个行人选取 4 张图像，组成容量为 512 的批次进行训练。经过骨干网络的基本信息后，在局部分支将特征图划分为 8 个水平部分，每个部分经过水平池化输出特征的维度为 3840 维向量。在全局分支进行困难样本挖掘，随后将采样索引传递给局部分支以减小后续进行局部特征融合时的计算量。然后将使用上述改进后的损失函数对模型进行优化。在模型训练和测试阶段在局部分支进行特征融合计算时λ_{att}设置为 12.0。其余设置于第三章保持一致。

4.4.2 实验结果及分析

为验证局部特征融合的切块对齐方法的有效性，分别在 Market1501 和 DukeMTMC-reID 数据集上进行训练。保持其他设置不变的情况下，在 MFSINet 局部分支加入局部特征融合的切块对齐方法得到切块对齐网络 SANet。实验结果汇总到表 4.1 所示。

由表 4.1 可知，SANet 相比于增强的基线Baseline^{+}拥有较强的提升，Rank-1 达到 95.4%，mAP 达到了 87.4%。该结果表明了在 Market1501 数据集上，融入本章算法的模型结构可以更有效的提取行人特征。

表 4.1 添切块对齐网络模型性能

Table 4.1 performance of slice alignment network model

Model	Market-1501		DukeMTMC-reID	
	rank-1	mAP	rank-1	mAP
Baseline^{+}	93.5	84.3	83.8	72.5
SANet	95.4	87.4	88.9	77.1

在 DukeMTMC-reID 数据集上对行人重识别率相较与增强的基线也有一定幅度的提升，在 Rank-1 上提升了 5.1%，mAP 指标提升了 4.2%。该结果表明了在 DukeMTMC-reID 数据集上，本章方法同样使有效的。

（1）消融实验

添加了基于局部特征融合的切块对齐方法后模型性能有了显著提升，为了验证本章所提出的方法对与第三章所提出方法在模型性能提升上的独立性，本章在 Market1501 和 DukeMTMC-reID 均进行了消融实验。

本章在两个数据集上分别基于增强基线和第三章模型同时添加了局部特征融合模块 LFFM 和双分支困难样本挖掘 DBHM，并进行了实验。此处消融实验主要证明本章所提出的方法可以独立于第三章的方法，对模型性能起到正向影响提升。且 DBHM 主要作用时帮助 LFFM 获得更多不对齐样本，所以两部分密不可分。所以此处选择将 LFFM 与 DBHM 组合在一起，对 SANet 中第三章所提出的改进方法进行消融对比。

表 4.2 Market1501 数据集上消融实验结果

Table 4.2 ablation experimental results on Market1501 dataset

Baseline+	+MRFM +SIFM	+LFFM +DBHM	Rank-1	mAP
√			93.5	84.3
√	√		94.8	86.5
√		√	94.5	86.6
√	√	√	95.4	87.4

由表 4.2 所示，在 Market1501 数据集上基于强力基线添加局部特征融合模块和双分支困难样本挖掘方法后 Rank-1 指标提升了 1.0%，而 mAP 指标提升了 2.3%。在基于第三章模型添加本章方法后 Rank-1 提升了 0.6%，mAP 提升了 0.9%。

由表 4.3 所示，在 DukeMTMC-reID 数据集上基于强力基线添加本章方法后 Rank-1 指标提升了 2.6%，而 mAP 指标提升了 2.7%。在第三章模型添加本章方

法后 Rank-1 提升了 2.3%，mAP 提升了 1.2%。

实验表明，无论是第三章还是本章所提出的方法，均能独立且有效地帮助模型提取出更据判别性的特征，提升模型对行人的识别准确率。

表 4.3 DukeMTMC-reID 数据集上消融实验结果

Table 4.3 ablation experimental results on DukeMTMC-reID dataset

Baseline+	+MRFM +SIFM	+LFFM +DBHM	Rank-1	mAP
√			83.8	72.5
√	√		86.6	75.9
√		√	86.4	75.2
√	√	√	88.9	77.1

（2）相关方法比较

本章提出的算法模型与近几年提出的算法进行了比较。本文中的方法总体上基于全局特征和局部特征相融合的思路，并在本章中对局部分支进行了改造和优化，并使用样本对切块之间的注意力对样本进行了空间上的重构使之对齐。本章选取了以下方法进行比较。首先，选择了预定义切分提取局部特征的方法分别进行了比较。然后，针对解决姿态视角差异的问题与基于软注意力机制的相关方法进行了比较。最后同样还与目前比较热门的基于生成对抗网络的方法进行了比较。

首先，本章方法与 Market-1501 数据集上的相关算法比较结果详见表 4.4。与 DukeMTMC-reID 数据集上的算法比较见表 4.5 所示。

表 4.4 Market-1501 数据集上的相关方法对比

Table 4.4 comparison of relevant methods on market-1501 dataset

方法	Rank 1	mAP
PGFA[17]	91.2	76.8
VPM[16]	93.0	80.8
HPM[14]	94.2	82.7-
OSNet[19]	94.8	84.9
CAM[21]	94.7	84.5
HOA[65]	95.1	85.0
AANet[66]	93.9	83.4
P2-Net[27]	95.2	85.6
IDE+UnityStyle[67]	93.2	89.3
MpRL[68]	85.8	67.5
DG-Net[71]	94.8	86.0
CAD-Net[69]	83.7	-
Ours（SANet）	**95.4**	**87.4**

由表 4-4 所示，在 Market-1501 数据集上经过本章算法优化后的模型能在相关方法比较中 Rank-1 指标处于领先位置。mAP 指标仅落后于基于图像风格迁移

的方法 IDE+UnityStyle，但 Rank-1 指标比该方法高出 2.2%。

表 4.5 DukeMTMC-reID 数据集上相关方法对比

Table 4.5 comparison of relevant methods on dukemtmc Reid dataset

方法	Rank 1	mAP
PGFA[17	82.6	65.5
VPM[16]	83.6	72.6
HPM[14]	86.6	74.3
OSNet[19]	88.6	73.5
CAM[21]	85.8	72.9
AANet[66]	87.7	74.3
PAT[60]	88.8	78.2
P2-Net[27]	86.5	73.1
IDE+UnityStyle[67]	82.1	65.2
MpRL[68]	78.8	58.6
DG-Net[71]	86.6	74.8
CAD-Net[69]	75.6	-
Ours	**88.9**	**77.1**

由表 4-5 所示，在 DukeMTMC-reID 数据集上，经过本章算法优化后的模型能在相关方法比较中 Rank-1 指标同样处于领先位置。mAP 指标仅落后于基于软注意力的 PAT 方法，但 Rank-1 指标仍高出该方法。

（3）参数敏感性实验

表 4.6 Market1501 数据集上关于λ_{att}参数敏感性实验结果

Table 4.6 experimental results of λ_{att} parameter sensitivity on Market1501 dataset

λ_{att}	Rank 1	mAP
1.0	88.1	78.8
2.0	91.2	81.2
4.0	93.1	85.5
8.0	95.1	87.1
12.0	**95.4**	**87.4**
16.0	95.2	87.3
32.0	95.2	87.3
64.0	95.2	87.2

本章基于局部特征融合的切块对齐方法中，最为重要的步骤就是在式 4.9 中通过局部特征距离矩阵估算切块间注意力。局部特征距离矩阵经过正则化后通过λ_{att}进行放缩，然后经过$softmax$函数进行概率估算。λ_{att}代表的缩放尺度直接影响最终切块相似度估算结果，所以本文对λ_{att}的取值在 Market1501 上进行了相关实验，并进行相关可视化说明。

本章对λ_{att}设置了 10 组数值并分别进行实验。由表 4.6 可知λ_{att}的选择对算法性能影响很大。若λ_{att}取值为 1、2、4 时，导致最终模型性能下降。当λ_{att}值不

小于 8 时，对模型最终效果起正向作用。在取值为 12 时能够达到最高性能提升。当数值超过 16 之后，模型性能趋于稳定，并保持在一个较高的水平。

为了解释参数λ_{att}对模型的影响，本文图 4.3 中非对齐正样本对进行特征提取，通过局部特征计算了不同λ_{att}值下的注意力矩阵。由于本文将图像分为 8 个特征块，一对样本得到一组 8*8 的注意力矩阵，然后根据注意力值大小生成伪色彩图像进行了可视化。图中左侧数字为锚点切块序号，下方为样本切块序号。

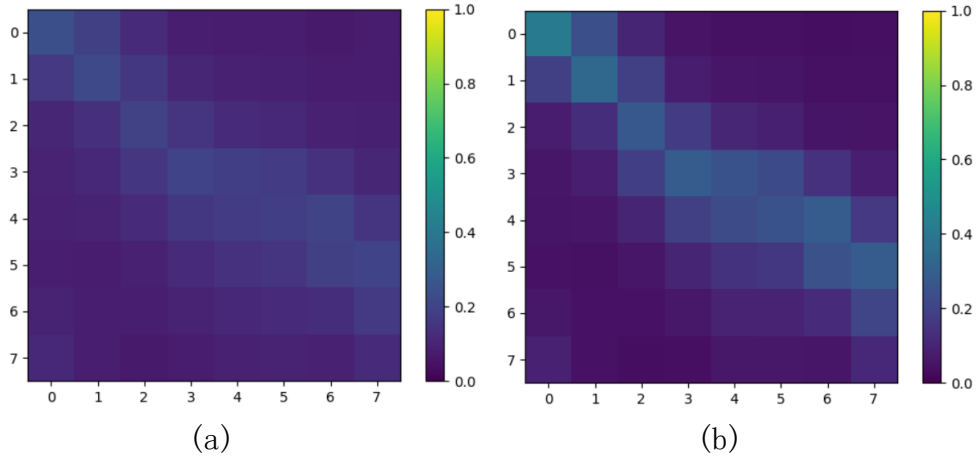


图 4.13 λ_{att}取 1.0(a)和 2.0(b)时注意力矩阵可视化

Fig 4.13 visualization of attention matrix when λ_{att}takes 1.0 (a) and 2.0 (b)

由图 4.13 中(a)可知，当λ_{att}取值 1.0 即原始值经过正则化后未进行尺度变换，图像趋近于被均匀混合，导致模型性能下降。由图 4.13 中(b)可知，当λ_{att}取值为 2.0 时，相较于 1.0 有一定幅度的提升，当时特征整体任然处于过度混合的状态。

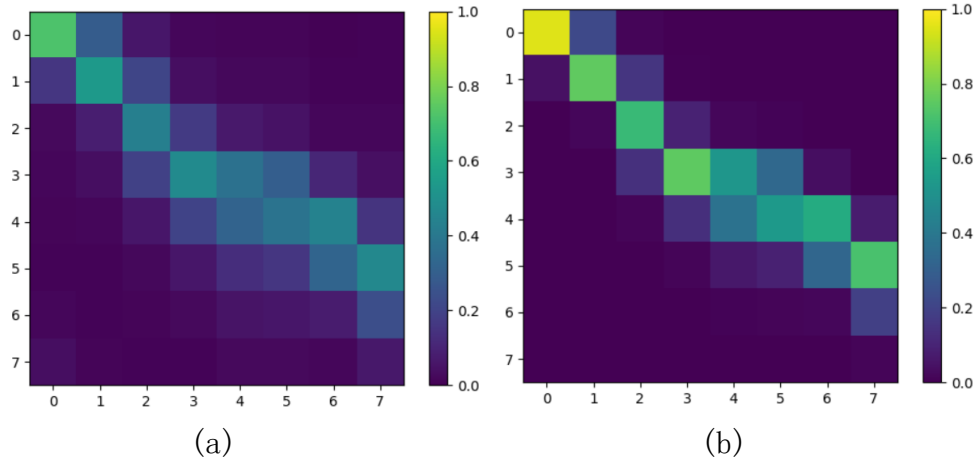


图 4.14 λ_{att}取 4.0(a)和 8.0(b)时注意力矩阵可视化

Fig 4.14 visualization of attention matrix when λ_{att}takes 4.0 (a) and 8.0 (b)

由图 4.14 中(a)可知，当λ_{att}取值为 4.0，模型倾向于将需要每个切块与 2 至 3 个切块进行关联，但是在一定程度上任然存在过度融合。由图 4.14 中(b)可知，

当λ_{att}取值为 8 时，模型能够关注集中在 1 至 2 个切块区域内，并且当切块间对齐时获得大于 80%的关注度。在不对齐的部分相似度很低，将关注度以大于 40%的比例集中在某连个相连接的部分，这于我们对图像的直观分析时一致的。

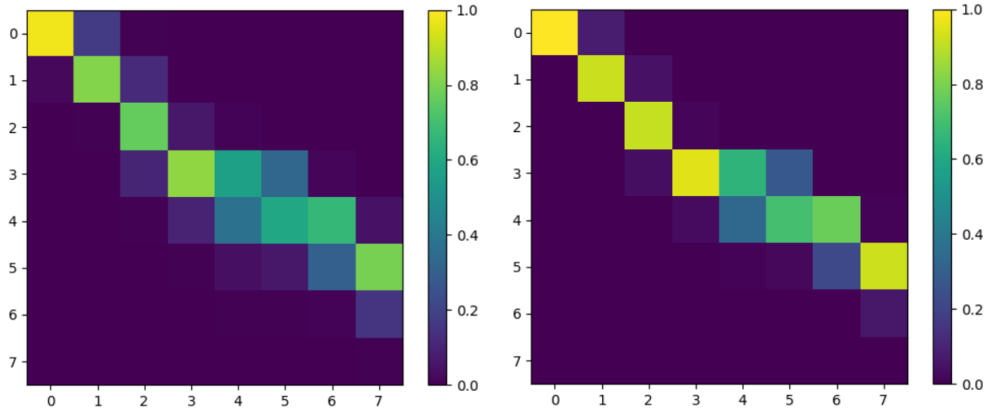


图 4.15 λ_{att}取 12.0(a)和 16.0(b)时注意力矩阵可视化

Fig 4.15 visualization of attention matrix when λ_{att}takes 12.0 (a) and 16.0 (b)

由图 4.15 中(a)可知，当λ_{att}取值为 12.0 时，模型能够准确的捕获切块间相似度，并且精准地根据距离计算出对应的注意力值。一般而言，行人图像在垂直方向上的拓扑结构，基于预定义大小的切块后行人结构并不能做到完全的分离。即便是将对齐的样本对以本章标准进行划分，相邻的其他部分都会存在部分重合信息。这一现象可能有行人姿态，动作等影响。所以注意力图像中，相邻图像存在一部分取值很小的关注是极为合理的。由图 4.15 中(b)可知，当λ_{att}取 16.0 时，模型京可能尝试使用一个最为相似的切块作为融合参照，而其他部分则趋近于不进行关注。

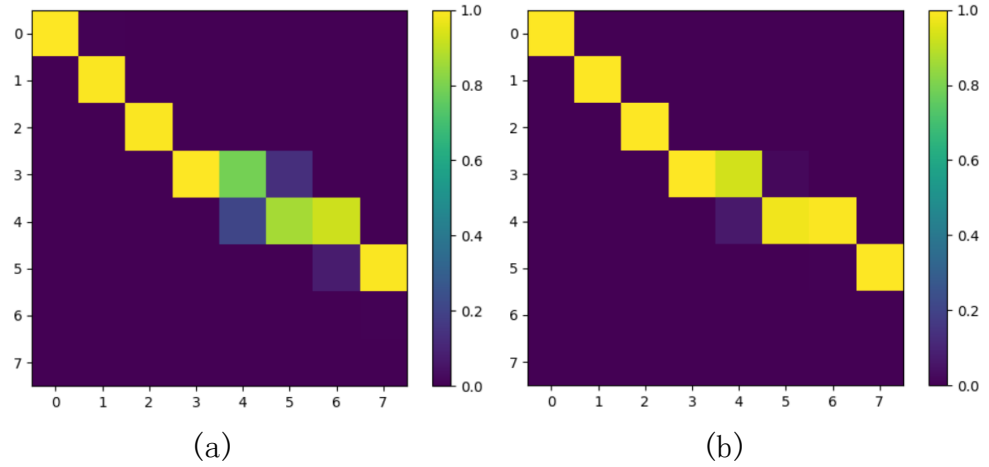


图 4.16 λ_{att}取 32.0(a)和 64.0(b)时注意力矩阵可视化

Fig 4.16 visualization of attention matrix when λ_{att}takes 32.0 (a) and 64.0 (b)

由图 4.16 中可知，当λ_{att}取值为 32.0 和 64.0 时，较当λ_{att}取 16.0 时的模型更为极端地将关注关注区域限制在某一最相似的区域。此时，模型在局部特征度量时等同于寻找图像最相近的切块进行比较，并计算每块切块距离并相加作为最终的距离度量。

综上所述，当λ_{att}取值为 12.0 时可获得最佳性能表现。取值过小将会导致局部特征过渡融合降低模型性能。较大的取值会将特征融合退化为寻找最相似特征块进行比较，在融合阶段倾向于舍弃相邻切块中的信息，但是同样对于解决图像样本不对齐起到正向作用。

（4）对齐算法可视化及原理分析

结合本章前文所分析的样本的特点，人体在垂直方向上的拓扑结构，大致可分为头部、上身、躯干、下肢等。本文以中的方法也水平切割的方式进行局部特征提取。但是，这种简单的均匀切分，并不总能在图像块得到均匀一致的人体结构。选择了四对样本分别为：对齐的正样本 ap，不对齐的正样本 up，对齐的负样本 an 和不对齐的负样本 un。其中对齐的正样本，和对齐的负样本在此问题上不会因人体的拓扑结构引入额外误差。本章提出了基于局部特征融合的切块对齐方法，减小这一误差对模型性能的影响。所以本小结将验证前文算法在对于区分 up 和 an 过程中起到的作用。

（i）实验现象:

以上分析可得，在模型进行切块对齐使，以局部距离计算出的注意力矩阵在四种类型的样本之间存在明显的差异。将其用于后续局部对齐后计算局部特征距离使带入的影响也随之不同。为此，本节将上述四组样本在对齐前后局部特征距离进行了度量，如图 4.7 并进行了汇总分析。

表 4.7 样本对齐前后局部特征距离变化表

Table 4.7 variation of local feature distance before and after sample alignment

	对齐正样本	不对齐正样本	对齐负样本	不对齐负样本
原始距离	1.68	2.12	2.09	2.46
对齐距离	1.55	1.66	1.88	2.14
变化值	-0.13	-0.46	-0.21	-0.32
变化率	-7.51%	-21.70%	-10.05%	-13.01%

通过表 4.7 中的数据，可以得到如下图所示为在使用和不使用局部特征融合的情况下，四组图像通过对局部特征距离按照升序排序的结果。图中表明，在不使用局部特征融合时，对齐的负样本位于不对齐的正样本之前，模型出现了错误判断。而使用局部特征融合之后，不对齐的正样本与锚点之间的距离缩进到对齐

的负样本之前，随意获得了正确的排序结果。其中不对齐样本对相较于对齐的样本对显然对于对齐算法更具敏感性。不对齐正样本对距离缩进绝对值为 0.398，百分比为 19.30%，均为 4 组样本中最高，从而有效地提升了正样本在候选集合排序中的相对位置。

如图 4.17 为使用本章方法前后对四组样本进行特征提取后，按局部特征距离进行升序排序的结果。图中蓝色框选出的为正样本，红色框选的为负样本。经过实验表明，本章中提出的基于切块的局部特征空间对齐方法是有效的。对齐的样本由于其人体拓扑结构相差不大，因此在空间重构的过程中趋于与原本一致，则期间距离缩进较小。而不对齐样本进行对齐计算后，在特征空间中的距离会大幅降低。而不对齐的负样本由于本身就位于排序末尾，其之间的距离减小对最近结果影响不大。而大幅拉近的不对齐正样本，则可以在排序结果中处于靠前位置，并最终被确定为候选结果。

图 4.17 不同类样本距离排序受对齐算法影响结果

Fig. 4.17 the results of different types of sample distance sorting affected by alignment algorithm

（ii）算法原理分析:

为了对实验结果结果进行原理分析，以下将四组样本在对齐算法计算过程中的切块间注意力矩阵进行可视化。在这一部分的可视化中，将以注意力数值大小生成伪色彩图像进行展示。并且在图像切块之间设定 0.25 为阈值，低于此阈值设为蓝色，高于此阈值设为红色，并设置图像连线宽度与注意力值成正比进行切块间对齐可视化。注意力矩阵可视化细节与上文实验中保持一致。

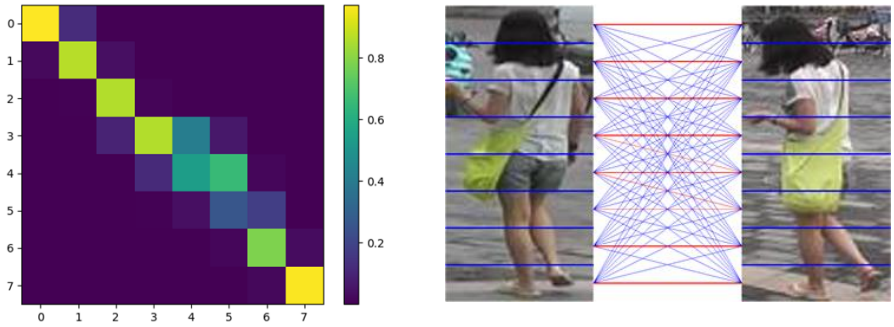


图 4.18 对齐的正样本对 ap 注意力矩阵及对齐可视化

Fig 4.18 visualization of attention matrix and alignment of aligned positive sample pairs

如图 4.18 为对齐的正样本对 ap，经过模型提取出局部特征进行相似度度量后，左侧各切块趋近于关注另一图像中水平与自身相同高度的切块。所以注意力矩阵也呈现出趋近于单位对角阵。

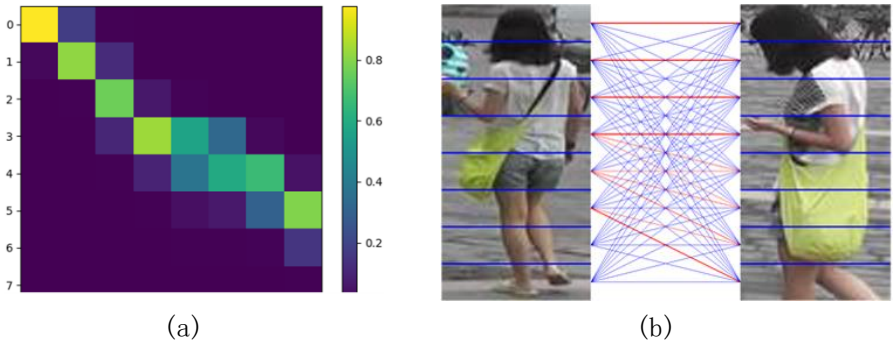


图 4.19 不对齐的正样本对 up 注意力矩阵及对齐可视化

Fig 4.19 visualization of attention matrix and alignment of unaligned positive sample pairs

如图 4.19 为不对齐的正样本对 up，由于右侧行人在图中占比较大，出现不对齐现象。左侧各切块趋近于关注另一图像中内容相近的部分。由于图像上半部分大致对齐，而下半段错位明显。所以，注意力矩阵也呈现出上半段以对角元素为主，下半段由 1 至两区域融合。

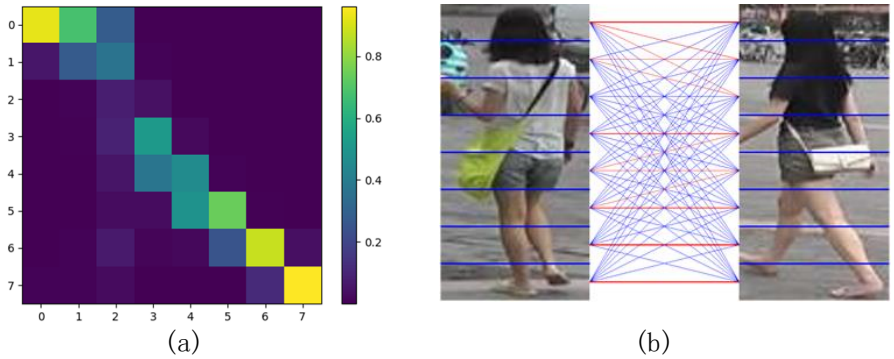


图 4.20 对齐的负样本对 an 注意力矩阵及对齐可视化

Fig 4.20 visualization of attention matrix and alignment of aligned negtive sample pairs

如图 4.20 为对齐的负样本对 an，由于图像中行人整体在空间上的拓扑结构

是对齐的，所以注意力矩阵中元素整体仍然呈现出对角线排列。但是由于局部切块在内容上是不同的，所以左边的切块并不能将注意力全部分配在右侧相同高度的切块中。所以在在具有明显差异的第 3 至 5 切块内，色彩图像中对应区域表现为多个暗色色块，表示切块的关注点存在明显的分散。

如图 4.21 为不对齐的负样本对 un，由于图像中行人整体在空间上的拓扑结构是不对齐的，且由于局部切块在内容上是不同的，切块的关注点存在明显的分散。但是能看出大致的趋势任然是以寻找对齐的拓扑结构为主。

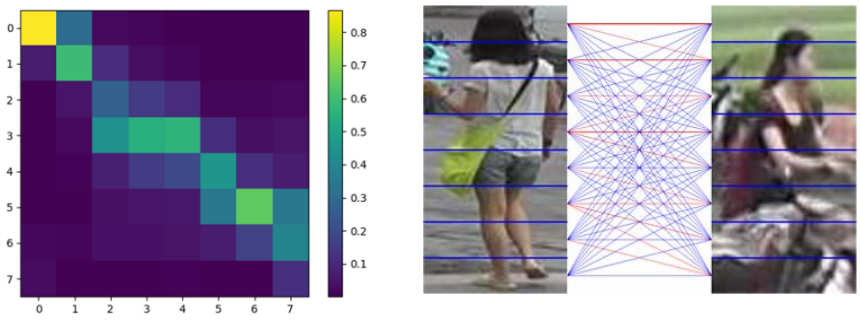


图 4.21 不对齐的负样本对 un 注意力矩阵及对齐可视化

Fig 4.19 visualization of attention matrix and alignment of unaligned negtive sample pairs

（iii）分析结论总结:

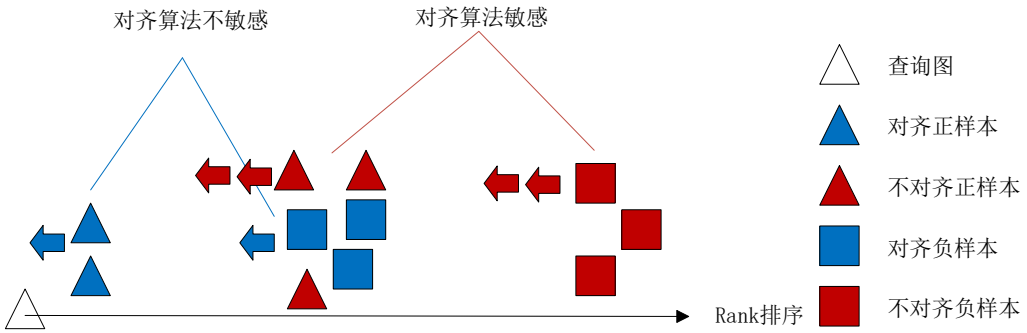


图 4.22 不同类样本对齐算法敏感性分析

Fig 4.22 sensitivity analysis of different types of sample about alignment algorithms

由上文分析可知，将基于局部特征融合切块对齐方法、在不同类型的样本之间呈现不同的特性。其中不对齐样本注意力矩阵会明显偏离对角线位置，所以在经过对齐算法进行优化后，在特征空间的距离明显降低。而对齐样本由于其本身水平切块处于对齐状态，所以基于局部特征距离计算除的对齐注意力矩阵，趋近于单位对角矩阵，即对角线元素为 1 其他元素为 0。那么根据注意力矩阵融合后与原始特征相似，会呈现出对齐样本对于对齐算法不敏感的现象。

如图 4.22 所示，四类样本可划分为对齐算法敏感型和对齐算法不敏感型。如图 4.22 和图 4.23 所示，本章算法原理就是利用不对齐正样本的对齐算法敏感

性这一特性，将重识别最终排序结果靠前且容易混淆对齐正样本和不对齐负样本进行了分离，以提升最终行人重识别准确率。

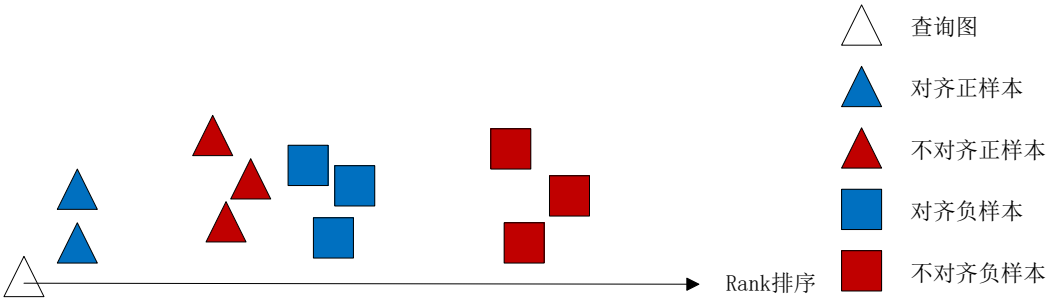


图 4.23 不同类样本受对齐算法影响结果

Fig 4.23 results of different types of samples affected by alignment algorithm

4.4 本章小结

本章基于局部特征融合的切块对齐方法。首先，介绍了目前流行的行人重识别数据集合现实场景中，行人图像在垂直方向上拓扑结构不对齐的问题。并正对这一问题提出了解决思路。然后，提出了基于全局分支和局部分支的网络模型结构，并对局部分支使用基于局部特征融合的切块对齐方法，解决提出的问题。最后，对提出的方法进行实现并在开源数据集上进行实验验证了本章方法的有效性。

第 5 章 行人重识别可视化系统设计与实现

基于前文的方法研究，已经能够得到具有识别率的行人重识别算法和基于 Pytorch 框架训练的网络模型。本章通过结合目前互联网开发领域的主流技术与上文研究成果，对行人重识别可视化系统设计与实现。

5.1 系统需求分析

本文已经通过对行人重识别方法的研究获得了具有识别率的行人重识别算法，通过算法训练出的网络模型可以应用于行人检索。对于整个流程中的关键步骤，本文进行了梳理。其中，首先通过上游行人检测获取行人重识别训练数据集。然后，训练行人重识别网络模型。接着，上传行人重识别候选集（Gallery）。最后，输入带查询样本（Query）模型计算并输出排序后的查询结果集。

若由用户直接运行上文行人重识别算法，在实际场景中存在许多缺点，如下：

（1）直接使用算法部分代码进行使用，需要用户具有较强的编码能力。

（2）由于需要更新候选图集图像库，需要部署机具有一定的存储能力；且直接由用户操作流程复杂且容易出错导致程序错误，可维护性和可运维性低下。

（3）本文算法基于学习进行研究并实现，在训练和使用过程中需要使用大量的 GPU 资源，直接部署用户机成本过高。

（4）行人重识别任务由于涉及到监控数据，具有一定的用户隐私敏感性，直接将算法暴露给使用者存在一定的安全风险。

综上原因，本文将通过结合目前主流技术设计并实现具有易用性的行人重识别系统。并且系统具有以下需求：

（1）用户登陆鉴权，对保护系统安全和数据安全。

（2）对行人重识别数据集进行展示。

（3）更新和管理不同的模型。

（4）输入查询图像，并选择数据集与模型执行重识别任务。

5.2 系统总体设计

由前文的需求分析可知，行人重识别系统最小实现单元需要具备以下功能：

首先，启动系统后进入登陆模块，用户需要先进行登陆才能进入系统。系统

初始设置又管理员账户。

进入系统后可以选择进入各子模块，其中数据集展示模块用于展示数据集信息。其中包括数据集统计信息，包括数据集图象数、相机数和训练集中行人类别数，用于帮助用户了解数据集总体信息。另外还从训练集、查询集与待查集三个方面展示数据集样图信息，用于帮助用户了解数据集的样图特点。

在模型管理模块可以对模型进行综合管理，其中包括增加和删除模型。由于模型均采用上文中算法中使用的结构，所以模型间的差异仅体现在模型参数不同，所以为了减小存储资源浪费，此处对模型的增删操作仅为模型参数文件的更新。

在重识别模块，需要先输入查询图并选择数据集与模型条件。系统将根据选择的参数文件加载模型，并对输入的查询图进行特征提取。然后根据选择的数据集对候选集中的图像进行特征提取。最后将查询图和候选集中的特征进行距离度量，并根据结果进行排序并返回查询结果。

登陆模块用于用户身份信息的登陆和鉴权管理，以保证系统和数据的安全性。

（1）数据集展示模块，用于管理和展示数据集。

（2）模型管理模块，用于模型集中管理。

（3）重识别模块，用于调用本文算法并隐藏操作细节，用于简化用户操作。

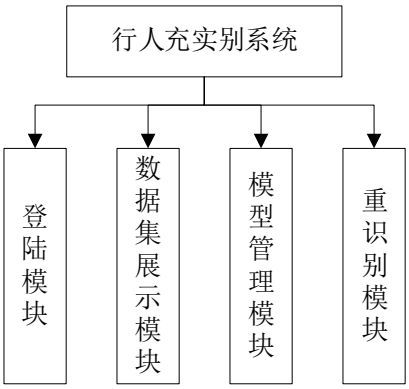


图 5.1 行人重识别可视化系统功能模块划分

Fig 5.1 functional module division of Person re-ID visualization system

5.3 系统实现与交互界面展示

如图 5.2 所示系统启动初始界面是登陆界面，用于用户身份信息的登陆和鉴权管理，以保证系统和数据的安全性。输入预设置的管理员账号后默认进入数据集展示界面。

如图 5.3 所示数据集展示页用于展示数据集信息。其中包括数据集统计信息，包括数据集图象数、相机数和训练集中行人类别数，用于帮助用户了解数据集总

体信息。另外还从训练集、查询集与待查集三个方面展示数据集样图信息，用于帮助用户了解数据集的样图特点。

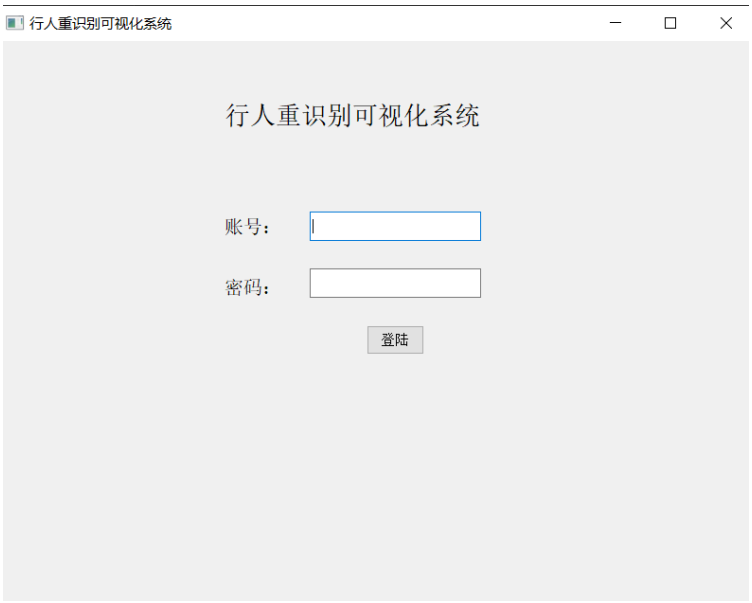


图 5.2 登陆界面演示图

Fig 5.2 demonstration of login interface

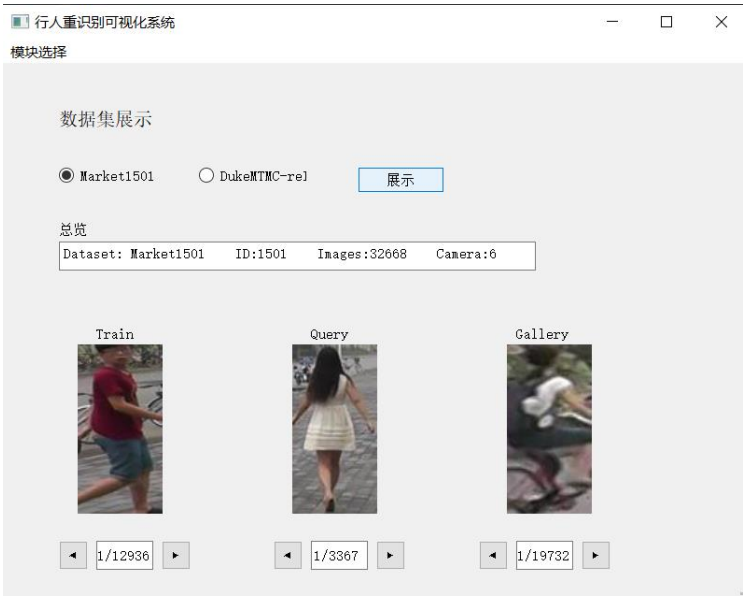


图 5.3 数据集展示界面演示图

Fig 5.3 dataset display interface

随后可以点击左上模块选择按钮进行页面切换，随后点击重识别模块进入如图 5.4 所示执行行人重识别任务主要的功能页面。系统默认将当前启动路径作为根路径，其下存在三个文件夹 queries、model_parameters 和 galleries 作为查询集、模型参数和候选集的根路径。在模型参数根路径下存放模型参数。在候选集根路径下存在多个子文件夹，其中存放各数据集的查询集图像集合。页面在进行初始

化时将遍历模型和候选集根目录，获取模型文件名和候选集根目录下各文件夹名最为后续模型和 gallery 选择的子选项。

图 5.4 行人识别页面演示图

Fig 5.4 demonstration of Person re-ID interface

在该页面中首先可以选择查询图，并进行展示。然后选择模型参数和待查询图像集所在目录，然后点击按钮执行检索操作。后台开始根据所选择的参数加载模型，然后进行特征提取并根据计算距离，并根据距离进行排序得到距离最近的样本最为识别结果并进行展示。

然后用户可以选择查询结果并点击分析按钮，系统将对查询图和结果图进行分块并进行局部块之间的注意力值，用于辅助用户进行后续进行更细粒度的人工判别。

5.4 本章小结

本章的主要是行人重识别系统的设计与开发。首先说明了系统需要实现的基本功能，在给出针对每个功能具体的设计实现思路，最后对整个系统进行各个模块的功能进行展示。

第 6 章 总结与展望

6.1 论文工作总结

在我们的日常生活中每天都会产生海量的监控视频，并且这些数据在安防和刑侦等领域有着巨大的利用价值。以过去人工进行比对的方式，在现如今的数据规模下想要快速且准确地进行行人识别十分困难。本文聚焦于这一流程中的行人重识别任务进行研究。该任务是在不同的时间、不同的摄像机下的行人图像集合中，给出一个特定行人并进行检索的问题。本文所作具体工作如下：

（1）结合近年来的研究并基于 ResNet50 提出一种基于全局特征融合的行人重识别网络结构，并进行优化得到一个强有力的增强基线。并且，针对现有的网络结构中不能利用底层及的卷积特征的问题进行了分析，并提出了基于多分辨率特征融合的方法。

（2）针对行人重识别中常见的样本背景复杂的问题，提出引入了空间信息融合的方法以解决背景信息干扰的问题。并将此方法融入网络模型中，提升模型在特征提取过程中对前景任务的关注，减弱背景信息的干扰。

（3）针对目前流行的行人重识别数据集和现实场景中，行人图像在垂直方向上拓扑结构不对齐的问题，并提出了基于局部特征融合的切块对齐方法，以减弱这一问题的干扰.

（4）结合本文的模型，针对最小功能集合实现了行人重识别可视化系统进行了设计与实现。该系统可以对本文算法结果进行清楚地展示。

6.2 进一步工作展望

本文基于深度学习提出了一个具有较高进度的行人重识别算法，并针对浅层卷积信息丢失、图像背景干扰和行人图像在空间内不对齐等问题提出了对应方法进行改进。但是本文提出的行人重识别方法仍有研究和继续改进的空间，具体如下：

（1）模型训练时局部分支进行特征融合的过程中，需要使用大量的内存资源。本文在实验过程中，对暂用内存较大的矩阵，进行了分块计算然后拼接的方式进行了优化，但是这一过程引入较大的时间复杂度。所以，后续可以在局部特

征提取阶段进行特征降维，减小内存的消耗。

（2）测试时在模型进行局部距离度量时，查询图像集与候选图像集中的行人图像需要一一进行对齐，同样会引入大量的计算量和时间复杂度。假定两图像集的数据规模为 m 和 n，每一次对齐计算时间为一个单位时间，则时间复杂度为 O(m*n)。后续可以通过选定特定的标准图像，将候选集对标准图像进行对齐，并将对齐后的局部特征进行持久化存储。在测试时只需要将查询集与标准图像进行对齐计算，然后将得到局部特征与存储的特征进行度量。则将复杂度降低为 O(m+n)。

（3）本文在测试阶段使用全局特征和局部特征分别计算距离矩阵，然后以加权求和的方式获得最后的结果作为行人判别依据，该过程中的对于计算结果受求和时权重影响较大。后续可以在距离度量之前将全局分支和局部分支的特征进行融合，测试阶段使用唯一的特征进行度量，使结果更加稳定。

参考文献

[1] 罗浩, 姜伟, 范星, 等. 基于深度学习的行人重识别研究进展[J]. 自动化学报, 2019, 45(11): 2032-2049.

[2] Zahra A, Perwaiz N, Shahzad M, et al. Person Re-identification: A Retrospective on Domain Specific Open Challenges and Future Trends[J]. arXiv preprint arXiv:2202.13121, 2022.

[3] Zajdel W, Zivkovic Z, Krose B J A. Keeping track of humans: Have I seen this person before?[C]//Proceedings of the 2005 IEEE international conference on robotics and automation. IEEE, 2005: 2081-2086.

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

[5] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets[J]. Advances in neural information processing systems, 2014, 27.

[6] Dalal N, Triggs B. Histograms of oriented gradients for human detection[C]//2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05). Ieee, 2005, 1: 886-893.

[7] Lowe D G. Object recognition from local scale-invariant features[C]//Proceedings of the seventh IEEE international conference on computer vision. Ieee, 1999, 2: 1150-1157.

[8] Koestinger M, Hirzer M, Wohlhart P, et al. Large scale metric learning from equivalence constraints[C]//2012 IEEE conference on computer vision and pattern recognition. IEEE, 2012: 2288-2295.

[9] Liao S, Hu Y, Zhu X, et al. Person re-identification by local maximal occurrence representation and metric learning[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 2197-2206.

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

[11] Wang F, Zuo W, Lin L, et al. Joint learning of single-image and cross-image representations for person re-identification[C]//Proceedings of the IEEE

Conference on Computer Vision and Pattern Recognition. 2016: 1288-1296.

[12] Qian X, Fu Y, Jiang Y G, et al. Multi-scale deep learning architectures for person re-identification[C]//Proceedings of the IEEE International Conference on Computer Vision. 2017: 5399-5408.

[13] Liu J, Zha Z J, Xie H, et al. Ca3net: Contextual-attentional attribute-appearance network for person re-identification[C]//Proceedings of the 26th ACM international conference on Multimedia. 2018: 737-745.

[14] 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.

[15] 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.

[16] Sun Y, Xu Q, Li Y, et al. Perceive where to focus: Learning visibility-aware part-level features for partial person re-identification[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 393-402.

[17] Miao J, Wu Y, Liu P, et al. Pose-guided feature alignment for occluded person re-identification[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2019: 542-551.

[18] Varior R R, Shuai B, Lu J, et al. A siamese long short-term memory architecture for human re-identification[C]//European conference on computer vision. Springer, Cham, 2016: 135-153.

[19] Zhou K, Yang Y, Cavallaro A, et al. Omni-scale feature learning for person re-identification[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 3702-3712.

[20] Li D, Chen X, Zhang Z, et al. Learning deep context-aware features over body and latent parts for person re-identification[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 384-393.

[21] Yang W, Huang H, Zhang Z, et al. Towards rich feature discovery with class activation maps augmentation for person re-identification[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 1389-

1398.

[22] Chen T, Ding S, Xie J, et al. Abd-net: Attentive but diverse person re-identification[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 8351-8361.

[23] Li W, Zhu X, Gong S. Harmonious attention network for person re-identification[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 2285-2294.

[24] 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.

[25] Zhao H, Tian M, Sun S, et al. Spindle net: Person re-identification with human body region guided feature decomposition and fusion[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 1077-1085.

[26] Zhao L, Li X, Zhuang Y, et al. Deeply-learned part-aligned representations for person re-identification[C]//Proceedings of the IEEE international conference on computer vision. 2017: 3219-3228.

[27] Guo J, Yuan Y, Huang L, et al. Beyond human parts: Dual part-aligned representations for person re-identification[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 3642-3651.

[28] Kalayeh M M, Basaran E, Gökmen M, et al. Human semantic parsing for person re-identification[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 1062-1071.

[29] 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.

[30] Wei L, Zhang S, Yao H, et al. Glad: Global-local-alignment descriptor for pedestrian retrieval[C]//Proceedings of the 25th ACM international conference on Multimedia. 2017: 420-428.

[31] Su C, Li J, Zhang S, et al. Pose-driven deep convolutional model for person re-identification[C]//Proceedings of the IEEE international conference on computer vision. 2017: 3960-3969.

[32] Liu J, Zha Z J, Tian Q I, et al. Multi-scale triplet cnn for person re-

identification[C]//Proceedings of the 24th ACM international conference on Multimedia. 2016: 192-196.

[33] Chen Y, Zhu X, Gong S. Person re-identification by deep learning multi-scale representations[C]//Proceedings of the IEEE international conference on computer vision workshops. 2017: 2590-2600.

[34] Zhou K, Yang Y, Cavallaro A, et al. Learning generalisable omni-scale representations for person re-identification[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021.

[35] Liu X, Zhao H, Tian M, et al. Hydraplus-net: Attentive deep features for pedestrian analysis[C]//Proceedings of the IEEE international conference on computer vision. 2017: 350-359.

[36] Ning X, Gong K, Li W, et al. Feature refinement and filter network for person re-identification[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2020, 31(9): 3391-3402.

[37] Deng W, Zheng L, Ye Q, et al. Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 994-1003.

[38] Huang Y, Zha Z J, Fu X, et al. Real-world person re-identification via degradation invariance learning[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 14084-14094.

[39] Chung D, Tahboub K, Delp E J. A two stream siamese convolutional neural network for person re-identification[C]//Proceedings of the IEEE international conference on computer vision. 2017: 1983-1991.

[40] Liu H, Jie Z, Jayashree K, et al. Video-based person re-identification with accumulative motion context[J]. IEEE transactions on circuits and systems for video technology, 2017, 28(10): 2788-2802.

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

[42] Yang J, Zheng W S, Yang Q, et al. Spatial-temporal graph convolutional network for video-based person re-identification[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 3289-3299.

[43] Wu Y, Bourahla O E F, Li X, et al. Adaptive graph representation learning for video person re-identification[J]. IEEE Transactions on Image Processing, 2020, 29: 8821-8830.

[44] Yan Y, Zhang Q, Ni B, et al. Learning context graph for person search[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 2158-2167.

[45] Shen Y, Li H, Yi S, et al. Person re-identification with deep similarity-guided graph neural network[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 486-504.

[46] Bai Z, Wang Z, Wang J, et al. Unsupervised multi-source domain adaptation for person re-identification[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 12914-12923.

[47] Liu J, Zha Z J, Wu W, et al. Spatial-temporal correlation and topology learning for person re-identification in videos[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 4370-4379.

[48] 郑远攀, 李广阳, 李晔. 深度学习在图像识别中的应用研究综述[J]. 计算机工程与应用, 2019, 55(12): 20-36.

[49] 李炳臻, 刘克, 顾佼佼, 等. 卷积神经网络研究综述[J]. 计算机时代, 2021, 4: 8-12.

[50] 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.

[51] Fan X, Jiang W, Luo H, et al. Spherereid: Deep hypersphere manifold embedding for person re-identification[J]. Journal of Visual Communication and Image Representation, 2019, 60: 51-58.

[52] Zhong Z, Zheng L, Kang G, et al. Random erasing data augmentation[C]//Proceedings of the AAAI conference on artificial intelligence. 2020, 34(07): 13001-13008.

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

[54] He K, Zhang X, Ren S, et al. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification[C]//Proceedings of the IEEE international conference on computer vision. 2015: 1026-1034.

[55] Lin G, Milan A, Shen C, et al. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 1925-1934.

[56] 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.

[57] Luo H, Jiang W, Zhang X, et al. Alignedreid++: Dynamically matching local information for person re-identification[J]. Pattern Recognition, 2019, 94: 53-61.

[58] 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-

[59] Zhang Z, Zhang H, Liu S. Person re-identification using heterogeneous local graph attention networks[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 12136-12145.

[60] 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.

[61] Gao H, Chen S, Zhang Z. Parts semantic segmentation aware representation learning for person re-identification[J]. Applied Sciences, 2019, 9(6): 1239.

[62] Lin M, Chen Q, Yan S. Network in network[J]. arXiv preprint arXiv:1312.4400, 2013.

[63] 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.

[64] Zheng Z, Zheng L, Yang Y. Unlabeled samples generated by gan improve the person re-identification baseline in vitro[C]//Proceedings of the IEEE international conference on computer vision. 2017: 3754-3762.

[65] Chen B, Deng W, Hu J. Mixed high-order attention network for person re-

identification[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2019: 371-381.

[66] 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.

[67] Liu C, Chang X, Shen Y D. Unity style transfer for person re-identification[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 6887-6896.

[68] Huang Y, Xu J, Wu Q, et al. Multi-pseudo regularized label for generated data in person re-identification[J]. IEEE Transactions on Image Processing, 2019, 28(3): 1391-1403.

[69] Li Y J, Chen Y C, Lin Y Y, et al. Recover and identify: A generative dual model for cross-resolution person re-identification[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 8090-8099.

[70] Chen X, Fu C, Zhao Y, et al. Salience-guided cascaded suppression network for person re-identification[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 3300-3310.

[71] Zheng Z, Yang X, Yu Z, et al. Joint discriminative and generative learning for person re-identification[C]//proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 2138-2147.

[72] Ming Z, Yang Y, Wei X, et al. Global-Local Dynamic Feature Alignment Network for Person Re-Identification[J]. arXiv preprint arXiv:2109.05759, 2021.

[73] Park H, Ham B. Relation network for person re-identification[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2020, 34(07): 11839-11847.

[74] Felzenszwalb P F, Girshick R B, McAllester D, et al. Object detection with discriminatively trained part-based models[J]. IEEE transactions on pattern analysis and machine intelligence, 2010, 32(9): 1627-1645.

攻读硕士学位期间的科研项目及获奖情况

科研项目：

项目名称：工业互联网边缘计算体系架构、协同优化与智能决策方法，自然基金联合基金（U1908212）

攻读硕士学位期间的专利和论文情况：

获得 2021~2022 东北大学硕士研究生学业奖学金