硕 士 学 位 论 文

双端共享网络的多模态行人重识别算法研究

导　师：

研究生：

东北大學

二〇二二年五月

分类号_____ 密级 _____

UDC _____

学 位 论 文

双端共享网络的多模态行人重识别算法研究

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： 工学

学科专业名称：计算机科学与技术

论文提交日期：2022 年 5 月 论文答辩日期： 2022 年 6 月

学位授予日期：2022 年 7 月 答辩委员会主席：

评 阅 人 ：

东 北 大 学

2022 年 月

A Thesis in Computer Software and Theory

Research on multimodal pedestrian re-recognition algorithm in Dual-terminal shared network

By

Supervisor:

Northeastern University

May 2022

独创性声明

本人声明，所呈交的学位论文是在导师的指导下完成的。论文中取得的研究成果除加以标注和致谢的地方外，不包含其他人已经发表或撰写过的研究成果，也不包括本人为获得其他学位而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示谢意。

学位论文作者签名：

日　　期：

学位论文版权使用授权书

本学位论文作者和指导教师完全了解东北大学有关保留、使用学位论文的规定：即学校有权保留并向国家有关部门或机构送交论文的复印件和磁盘，允许论文被查阅和借阅。本人同意东北大学可以将学位论文的全部或部分内容编入有关数据库进行检索、交流。

作者和导师同意网上交流的时间为作者获得学位后：

半年 □　　一年□　　一年半□　　两年□

学位论文作者签名：　　导师签名：

签字日期：　　签字日期：

摘 要

行人重识别的主要任务是在一段含有行人的视频或者图像中去寻找某个特定身份的行人，行人重识别任务可看作是图像检索问题。行人重识别的图像大多来源于监控图像，由于监控摄像头不仅会捕捉可见光图像，红外摄像头也会捕获红外图像，所以多模态下的行人重识别开始成为该领域中的研究热点。不同于单模态下的行人重识别，由于图像是多模态的，所以图像在识别中会因为存在模态差异而导致识别准确率降低。此外，图像的模态不同，所提取的特征也不同，如果使用同一个网络对不同模态的图像进行特征提取，那么网络中的参数需要能够对相应的模态信息有自适应性。在真实的场景下，由于图像都来自真实摄像头，经过行人检测算法得到的行人图像会因为摄像头的拍摄角度、拍摄距离和行人位置等问题，存在一定的姿态不对齐问题。这些样本对于行人重识别网络来说是难样本，网络模型很难学习到对姿态不对齐的样本的检测能力，只能对姿态正常的图像进行正确识别，而行人姿态不对齐的图像又是不可避免的，因此行人重识别网络对于正常图像存在一定的过拟合。在多模态行人重识别领域中，双端共享网络是一种比较常用的网络模型，本文基于双端共享网络研究多模态下的行人重识别算法，分别从表征学习和度量学习两方面对双端共享网络模型进行了优化，提出了使双端共享网络模型的重识别效果更好的优化方案。本文的具体工作如下：

（1）基于行人姿态不对齐的行人重识别数据增强方法。在真实场景下的行人重识别任务中，行人重识别网络的输入数据来源于行人检测算法从监控图像中获取的行人图像，其图像存在行人姿态不对齐的问题，而在训练过程中没有使用相关的图像对行人重识别网络进行训练，所以对姿态对齐的图像存在一定的过拟合。本文提出了一种基于行人姿态不对齐的数据增强方法，在训练之前将一部分数据进行收缩和填充，模仿真实场景下行人姿态不对齐的场景，将处理后的图像与正常图像一起作为训练数据输入网络中进行训练。使用多模态行人重识别领域中不同结构的代表性网络进行实验，可以有效降低网络模型的过拟合现象。

（2）基于非局部注意力机制的多模态行人重识别模型。在双端可共享的多模态行人重识别模型中，双端网络会分别提取特定模态下的特征，在进入共享特征提取网络之前，会对两种特定模态的特征进行融合，然后输入到共享网络中进行共享特征的提取。同一类的两种模态的特征内一定会存在相关联的共享特征，但在融合后的特征内，由于相关联的共享特征会存在较大的空间距离，且共享网络中卷积核的感受野大小有限，因此对于这种长距离的依赖共享网络无法有效的捕获。使用非局部注意力网络嵌入到共享特征提取网络中,可以将特征提取网络的感受野扩大到整个特征图，从而有效的提取到存在较大空间距离的共享特征。本文使用非局部注意力网络对 MACE 和 cm-SSFT 两种双端可共享网络模型进行优化，并进行了实验，实验表明，非局部注意力网络的加入可以有效提高行人重识别的准确率。

（3）基于特征均聚类损失的多模态行人重识别网络。行人重识别任务中一般使用三元组损失进行度量学习，为了提高训练效率，大部分网络模型都会进行难样本挖掘。但使用难样本挖掘的三元组损失只会对基准样本与难正样本和难负样本之间的距离进行优化，使得样本空间中仍然存在较大的类内距离和较小的类间距离，因此三元组损失对于多模态行人重识别任务并不理想。本文使用基于特征均值的聚类损失代替三元组损失进行度量学习，由于在优化过程中是基于特征均值来对特征距离进行优化的，所以间接的使得样本空间中所有的样本都参与了优化，同时也清除了较大的类内距离和较小的类间距离，使之能够将相同标签的样本进行聚类，从而使得不同类之间的样本更有区分度。本文使用基于特征均值的聚类损失函数分别对 MACE 和 cm-SSFT 两种双端可共享网络模型进行优化，并使用优化后的网络模型进行了实验，实验结果证明了该聚类损失函数有效提高了行人重识别网络的识别准确率。

关键词：多模态行人重识别；数据增强；卷积神经网络；聚类损失函数

Abstract

The main task of pedestrian re-identification is to search for pedestrians with specific identities in a video or image containing pedestrians, which can be regarded as an image retrieval problem. Most images of pedestrian re-recognition come from surveillance images. Since surveillance cameras capture not only visible images but also infrared images, multi-modal pedestrian re-recognition has become a research hotspot in this field. Different from pedestrian re-recognition under single mode, since the image is multi-mode, the recognition accuracy will be reduced due to modal differences in the image recognition. In addition, the modal of the image is different, the extracted features are different, if you are using the same network for different mode of image feature extraction, the parameters need to be able to in the network is adaptive to the corresponding modal information In a real scenario, because the images are from real camera, the pedestrian detection algorithm to obtain the pedestrian images will be because of camera shooting Angle Shooting distance and pedestrian position problems, there is a certain attitude alignment problem. These samples are difficult for pedestrian heavy recognition network samples, the network model is difficult to learn the posture is not aligned sample detection ability, can only to correct posture normal image recognition, and pedestrian posture is not aligned images is inevitable, so pedestrians have to identify network there is a fitting for normal images. In the field of the multimodal heavy pedestrian recognition, double-terminal shared network is a network model, which are frequently used in this paper, based on double side sharing network is studied under the mode of pedestrian recognition algorithm, respectively from two aspects: the characterization and measurement study of double-terminal shared network model is optimized, puts forward the heavy recognition effect is double-terminal shared network model better optimization scheme The specific work of this paper is as follows:

(1) A pedestrian re-recognition data enhancement method based on pedestrian posture misalignment. In real scenarios of pedestrian recognition task of heavy, heavy pedestrian recognition network's input data from the pedestrian detection algorithm from the monitoring image of pedestrians, the problems the pedestrian posture is not aligned images, and in the process of training is not related to the use of the image to pedestrian heavy recognition network training, so the posture alignment image has certain fitting. This paper proposes a data enhancement method based on the misalignment of pedestrian posture. Before training, part of the data is shrunk and filled to imitate the misalignment of pedestrian posture in real scenes, and the processed images and normal images are input into the network as training data for training Using representative networks of different structures in the field of multimodal pedestrian re-recognition can effectively reduce the over-fitting phenomenon of network model.

(2) Multi-modal pedestrian re-recognition model based on non-local attention mechanism. In the double-terminal multi-mode pedestrian re-recognition model, the double-terminal network will extract the features of specific modes respectively. Before entering the shared feature extraction network, the features of the two specific modes will be fused and then input into the shared network for shared feature extraction. There must be associated shared features in the features of the two modes in the same class. The non-local attention network embedded in the shared feature extraction network can expand the receptive field of the feature extraction network to the whole feature graph, so as to effectively extract the shared features with a large spatial distance. In this paper, non-local attention networks are used to optimize MACE and CM-SSFT double-terminal shared network models, and experiments are carried out. The experiment shows that the addition of non-local attention networks can effectively improve the accuracy of pedestrian recognition.

(3) Multimodal pedestrian reidentification network based on feature mean clustering loss. Triplet loss is generally used to measure learning in pedestrian

re-recognition tasks. In order to improve training efficiency, most network models will carry out difficult sample mining. However, triplet loss using difficult sample mining can only optimize the distance between the reference sample and the difficult sample and the difficult sample, so there are still large intra-class distance and small inter-class distance in the sample space. Therefore, triplet loss is not ideal for multimodal pedestrian re-recognition tasks. Used in this article, based on the characteristics of the average clustering loss instead of a triple loss measurement study, because in the process of optimization is to optimize the characteristic distance based on the characteristics of mean, so the indirect making all the samples in the sample space are involved in the optimization, also remove the large distance and the distance between the smaller class in class, he was able to turn the same tag clustering samples, from Which makes samples from different classes more differentiated. In this paper, the clustering loss function based on feature mean is used to optimize the MACE and cm-SSFT double-terminal shared network models respectively. The experimental results prove that the clustering loss function can effectively improve the recognition accuracy of pedestrian re-identification network.

Key words: Multimodal pedestrian re-recognition; Data Augmentation; Convolutional neural network; Cluster loss

目 录

第 1 章 绪 论

1.1 研究背景

行人重识别是利用计算机视觉技术对跨设备下的图像或视频进行匹配，即给定一个行人图像，在不同设备的图像库中检索出同一个行人。近年来，随着网络基础设施的建设，监控系统也越来越普及。当前的监控系统正在向智能化监控迈进，且朝着城市级应用发展，例如政府提出的“平安城市”、“智慧城市”以及“雪亮工程”等。应用范围的扩大导致监控数据也越来越大，而想要在海量数据中搜索某一目标也越来越难。尤其是在搜索某一个行人时，由于目标人物出现的时间、地点不是唯一的，所以搜索难度很大。而行人重识别的任务就是给定一个目标人物的图像，在已有的图像集中选出与目标人物身份一致的图像。当前的行人重识别领域面临着一系列难点，比如行人图像分辨率过低、行人身体部位被遮挡、监控环境的变化等，这些问题导致了行人识别的准确率下降。单模态下的行人重识别可以大致分为基于度量学习的行人重识别和基于深度学习的行人重识别。基于度量学习的方法是将提取到的各类行人特征投影到一个度量空间，并使得相同标签的行人特征距离更近，不同标签的行人特征距离更远。基于深度学习的方法是目前主流的方法，随着深度神经网络的高速发展，该类方法诞生了许多经典的框架，比如基于局部特征提取的框架、基于语义分割的框架、基于 GAN 的框架等。目前传统的单模态行人重识别算法在主流数据集上取得了较高的准确率。

当前传统的行人重识别领域借助于深度神经网络，在主流数据集上取得了较高的准确率，但大多数算法都是基于单模态图像的数据。在实际的监控系统中，为了实现全天候监控，一般会设置多种模态类型的摄像头，例如白天使用普通光学摄像头用于捕获可见光图像，夜晚使用红外摄像头用于捕获红外图像，由于两种图像的摄取方式不同，所以导致了两种图像之间存在模态差异，而传统的单模态行人重识别方法中学习到的神经网络只适用于提取可见光模态的特征，如果将图像换为和可见光图像存在模态差异的红外图像，则没有明显的效果。因此可见光图像的行人重识别网络并不适用于多模态的行人重识别问题。在当前的多模态行人重识别问题的

研究中，可大致分为两类，一类是借助与对抗生成网络（GAN）生成另一种模态图像来辅助训练神经网络，该类方法由于涉及到对抗生成网络的训练，使得训练成本额外增加，不利于模型的部署。另一类是使用双端网络提取不同模态的特征后再进行特征融合，然后进行分类。该类方法是将深度学习与度量学习结合的方法，特征提取模块采用双端网络分别提取不同模态的图像特征，并在全连接层上共享参数，得到跨模态共享特征，最后使用一个损失函数约束不同模态特征的距离，使得相同标签的各模态特征距离更小，不同标签的各模态特征距离更大。

1.2 研究现状

随着监控系统的普及，监控图像对于搜索目标人物起到了至关重要的作用，如何在大量的监控图像中高效的锁定到目标人物成为了重点。使用深度学习技术可以有效的解决这个问题。目前学术界对于单模态和多模态的行人重识别进行了大量的研究，提出了许多有效的方法。

1.2.1 行人重识别基本流程

行人重识别的任务是利用计算机视觉技术去判断图像或视频序列中是否存在特定身份行人的技术[1]，即：给定一个监控行人图像，去检索跨设备下的图像中改行人的图像，其本质上是一种图像检索的子问题。行人重识别类似于人脸识别，在监控图像中，由于拍摄距离和角度以及相机分别率的问题，通常无法得到人脸比较清晰的图像，人脸识别技术无法有效识别行人。行人重识别技术通过学习人体姿态、外形轮廓等信息，从给定的图像中检索特定身份的行人，对于智能安防、人机交互等领域有重要的意义。真实场景下行人重识别的数据来源于行人检测算法的检测结果，行人重识别系统如图 1.1 所示，其过程包括两个阶段：行人检测阶段和行人重识别阶段。行人检测阶段主要使用行人检测算法对原始视频帧进行行人检测，并将检测到的所有行人进行提取，用于行人重识别的输入数据。行人重识别阶段将检测到的行人数据作为 Gallery，并将待检索的目标行人图像作为 Query，通过行人重识别网络中学习到的参数对目标进行检索。由于行人重识别还有许多技术性问题有待研究，因此人们将行人检测阶段和行人重识别阶段放到两个框架下来研究。一般学术界研究的行人重识别算法

也主要是指行人重识别阶段使用的算法。

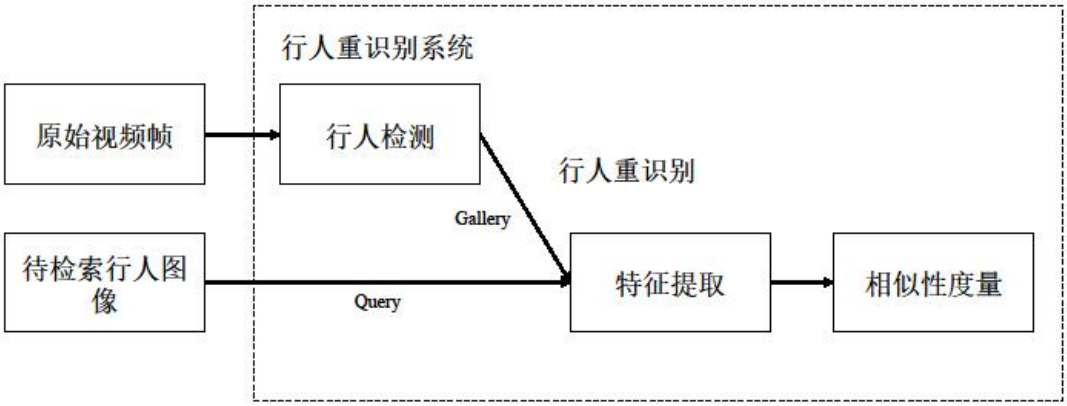


图 1.1 行人重识别系统示意图

Fig 1.1 Schematic diagram of pedestrian re-identification system

对于行人重识别网络的训练，其数据集除了包括训练集和验证集外，还包括 Query 集和 Gallery 集。如图 1.2 所示，行人重识别算法从 Query 集中得到一个关心的行人图像，以这个行人图像作为 Query，再从包含多张行人图像的 Gallery 集检索与该 Query 图像是同一身份的图像，并根据对每张图像的预测概率从大到小对 Gallery 集的图像进行排序，最后输出结果序列。

图 1.2 Query 和 Gallery 示意图

Fig 1.2 Query and Gallery schematic diagram

1.2.2 单模态下的行人重识别

单模态下的行人重识别任务主要由三个部分组成，分别是特征学习、度量学习和排序优化。特征学习旨在提取行人图像的全局特征或局部特征，如 Zheng 等人[2-4]提出的身份识别嵌入模型通过提取人物的全局特征，将每个标识视为一个不同的类，将训练过程构建为一个多类分类问题。但考虑全局特征的算法无法准确识别姿态不对齐的样本，Suh 和 Zhao 等人[5,6]提出结合全局特征与局部特征来对人物进行识别，此外，Cheng 等人[7]提出的多通道融合、Li 等人[8]提出的多尺度上下文感知卷积、Zhao 等人[9]提出的多级特征分解以及 Suh 等人[10]的双线性池化都是为了提高局部特征学习。对于水平分割的区域特征，Sun 等人[11]提出的 PCB 模型在解决姿态不对齐问题上也有出色的表现，该模型将不同部位通过部件分类器进行分类，以增强局部特征的学习。除了全局特征和局部特征，也可以利用一些辅助信息来增强特征学习的效果，包括附加的注释信息和增强训练样本，如 Su 等人[12]提出的一种融合了语义属性信息的深度属性学习框架，Chang、Liu 和 Zhu 等人[13-15]则使用融合了视觉角度的信息来辅助特征学习，Liu 等人[16,17]使用 GAN 生成不同姿态、不同背景的图像来强化特征学习。度量学习主要体现在损失函数的设计上，一般通过计算交叉熵损失函数来作为分类损失[2]，使用对比损失函数[18]或二分类损失函数[19]当作验证损失，以及使用三元组损失函数[20]来控制样本特征之间的距离。排序优化是根据图片之间的相似性对出事的结果进行优化，如 Ye 等人[21,22]提出的重排序和排序融合，都是将特征相似的正样本排名靠前从而使检索结果更加精确。

1.2.3 多模态下的行人重识别

多模态下行人重识别的数据可以有多种形式，有可见光-红外图像、图像-文本、跨分辨率图像等。其中可见光-红外图像的跨模态行人重识别吸引了学术界及工业界的广泛关注。在过去的几年中，学术界提出了大量的深度神经网络来解决跨模态下的行人重识别问题，这些网络在多个方面取得了一定的效果。在常用的模型中，根据主干网络的整体结构大致可分为三种类型，分别是单流结构（One-stream structure）、双流结构（Two-stream structure）和非对称全连接层结构（Asymmetric FC Layer Structure），三种

结构如图 1.3 所示。

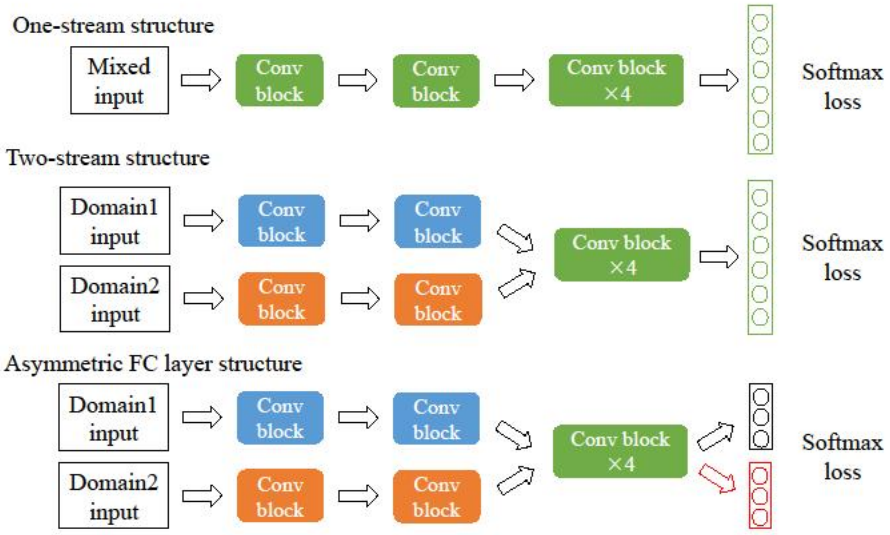


图 1.3 多模态行人重识别网络的三种结构

Fig 1.3 Three structures of multimodal pedestrian re-identification network

单流网络结构对于计算机视觉领域是应用比较广泛的一种结构，比较有代表性的网络包括 AlexNet[23]、 VGG[24]、GoogleNet[25]、ResNet[26]等，在行人重识别领域中用于单模态下的识别人物效果显著，对于多模态的场景，其效果并不理想。Wu 等人[27]提出一种使用深度零填充的单流神经网络的算法，如图 1.4 所示。

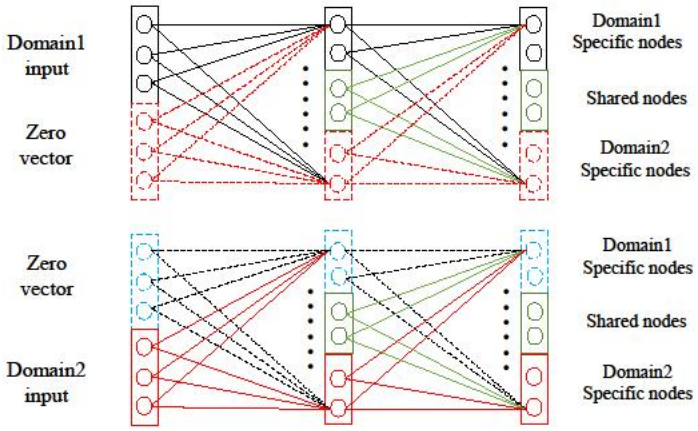


图 1.4 深度零填充网络

Fig 1.4 Deep zero-fill network

该算法将单一模态的输入进行零填充，使得所有输入都可作为单流网络的输入。以图 1.4 为例，在每一层中，蓝色节点表示域 1 的特定节点，红色节点表示域 2 的特定节点，绿色节点表示共享节点，虚线节点表示零

值。将填充后的数据输入网络后，网络中每层会根据输入的数据选择合适的域来进行响应，从而训练出具有特定域选择能力的神经网络。

双流网络结构适用于多模态的匹配任务，也算多模态行人重识别领域中使用比较多的网络结构。如图 1.3 所示，双流网络结构有两个输入端口，分别输入两个不同模态的数据。在较浅的网络层中，其参数是特定于某个模态的数据，在较深的网络层中，其参数是两个模态共享的。这样设定的主要思路是认为网络之所以能对不同模态的图像进行识别，是因为不同模态的图像具有某些共有的特征，只有识别了这些特征，才可以找到唯一的匹配项。依据这一理论，学术界提出了许多基于双流网络结构的算法，如 Ye 等人[28,29]提出的模态感知协作算法（MACE）、多模态双向中心约束算法（BDTR）等。此外，也有学者关注共享特征的同时，也关注了不同模态的特定特征，他们认为不同模态的特定特征也可以帮助网络识别图像，如多模态图像中 RGB 图像的色彩信息和红外图像的热力学信息都可以作为辅助信息来帮助网络进行目标匹配。基于这一理念，Wang 等人[30]提出的减少双极差异学习算法（D^2RL）、Yan 等人[31]提出的跨模态特征转移算法（cm-SSFT）等。这些算法的神经网络都是基于双流网络来关注特定模态信息的，对于多模态行人重识别的准确率都有不错的表现。在使用卷积神经网络（CNN）提取特征的同时，也有学者使用生成对抗网络（GAN）技术来处理不同模态之间的特征，如 Wang 等人[32]提出的模型中，在使用双流网络进行特征提取的同时，使用 GAN 技术生成不同模态行人的图像来减少不同模态间的特征差异。

非对称全连接层网络同样可用于跨域选择任务，由图 1.3 可知，该结构前面的网络层共享所有的参数，只有最后一层的全连接层是不共享的，这样设计的原因是假设了不同模态的特征进行提取时的过程是相同的，并且全连接层可对提取的特征进行模态的自适应。Chen 等人[33]提出的一致性正则化的非对称距离模型（CVDCA），该方法建立了一个非对称模型来学习特定的投影，并将每个视图的不匹配特征转化为一个公共空间，在这个公共空间中提取特定特征，同时对不同视角的特征变换进行相关性建模，避免使其发生过拟合。最初是用来解决单模态下图像与图像之间背景特征差别过大导致的识别力下降的问题，由于其应用场景符合跨模态图像的场景，所以也被用作多模态下的行人重识别解决方案。

单模态下的行人重识别技术已经达到相当高的准确率，现有的跨模态行人重识别技术在公共数据集上也取得了令人满意的结果，但在准确率以及网络结构的优化上还有许多不足。由于当前多模态摄像头应用广泛，跨模态下的行人重识别技术也有相当重要的应用基础，该技术对于理论研究和实际应用价值也都非常有意义。

1.3 本文主要研究内容

本文主要研究针对可见光-红外图像下的跨模态行人重识别技术，结合当前该领域的主流技术，总结了部分存在的主要问题，总结如下：

（1）由于实际应用中行人图像来自真实摄像头采集，导致存在大量姿态不对齐问题，而当前跨模态行人数据集中往往没有考虑到姿态不对齐的问题，导致训练所得到的网络对于真实场景下的数据并不具有鲁棒性。

（2）当前主流的双端网络结构可以提取不同模态图像的特征，并将跨模态特征进行融合，以进一步提取共享特征。但由于来自两张不同的图像，所以对于长距离关联的共享特征无法有效地提取，从而导致行人重识别的准确率无法提高。

（3）行人重识别使用度量学习来控制特征之间的距离，目前广泛应用的三元组损失在单模态下可以很好地控制特征距离，使用难样本挖掘的三元组损失只会对基准样本与难正样本和难负样本之间的距离进行优化，使得样本空间中仍然存在较大的类内距离和较小的类间距离，因此三元组损失对于多模态行人重识别任务并不理想针对以上问题，本文提出了以下的解决方案：

（1）针对真实数据中行人姿态不对齐、比例不一致等问题，提出一种针对训练数据的数据增强方法，使得数据更接近真实场景下的图像。该方法先将图像缩小比例，再对图像的上左右或下左右三边进行填充，使得图像与真实场景中行人姿态不对齐的场景一致。将处理后的数据与原数据一起作为训练数据作为输入，可以使训练后所得到的网络对真实数据的识别准确率更高，更具有鲁棒性。

（2）在共享网络中加入非局部注意力块（Non-local Attention Block），用来获得所在位置的加权和，非局部关注块可以突出重点关心的区域，消除噪声，从而汇聚更多有用的信息。传统的卷积神经网络（CNN）感受野

太小，而局部注意力块可以扩大感受野，将更大范围内与当前样本点有关联的联系起来，进而能够捕获长距离依赖，使得特征提取的效率以及准确率更高。

（3）使用聚类损失代替三元组损失。聚类损失基于均值来计算距离，使得损失函数不仅最小化难样本之间的距离，还间接地最小化所有类内图像之间基于均值的距离，从而使训练批次中的难样本对损失函数有直接贡献。

1.4 论文结构

本文的主要内容总共分为六个章节，具体如下：

第一章首先介绍了行人重识别的定义以及分类，然后介绍了当前各个分支领域的主要研究内容及发展状况，并指出了当前跨模态行人重识别领域存在的一些问题，最后介绍了本文的主要研究内容。

第二章主要介绍本文中涉及到的主要技术，首先介绍了深度学习中的卷积神经网络，分别介绍了其结构和扩展模型。之后介绍了深度学习中经常使用的一些特征融合算法。最后介绍了深度学习中常用的一些技术，包括常用的数据增强方法、激活函数、优化算法以及损失函数。这些技术也是多模态行人重识别技术的基础。

第三章主要针对当前数据集存在的问题进行了分析，指出了当前数据集没有充分包含真实数据中行人姿态不对齐的问题，并进一步分析了由此引发的识别网络对于数据缺乏鲁棒性的问题。针对这些问题，进而提出了一种数据增强方法，来保证数据集能够更好的模拟真实数据。

第四章分析了当前该领域网络模型在特征提取方面的不足，即无法获取特征融合后的长距离依赖。本文提出了使用非局部注意力技术来提高特征提取的效率，使得特征提取网络可以捕获长距离依赖的特征，并使用非局部注意力网络分别对两种双端网络进行了优化，最后进行了实验。

第五章针对模型的损失函数进行了分析，指出了当前使用的三元组损失在多模态行人重识别领域的不足之处，提出了一种基于特征均值的聚类损失函数，该函数使得每一个样本对网络训练都有贡献，有效提高了网络的训练效率和识别准确率。同时使用聚类损失函数对两种双端网络进行了优化，并进行了实验分析。

第六章首先总结了论文的工作，并介绍了该领域未来的研究方向。

第 2 章 相关技术概述

2.1 卷积神经网络

2.1.1 整体结构

卷积神经网络（CNN）是深度学习领域中最常见的一种网络结构，在计算机视觉中主要用于对图像进行特征提取。一般的图像本身由像素组成，如果直接输入网络中参与运算，将会导致参数太多，使得计算效率降低。卷积神经网络通过提取图像的特征，将问题进行简化，保留图像主要特征的同时也将神经网络中参数大量减少，并且不会影响网络对图像的处理结果。卷积神经网络主要由三部分构成，分别是卷积层、池化层和全连接层。常见的卷积神经网络结构如图 2.1 所示。

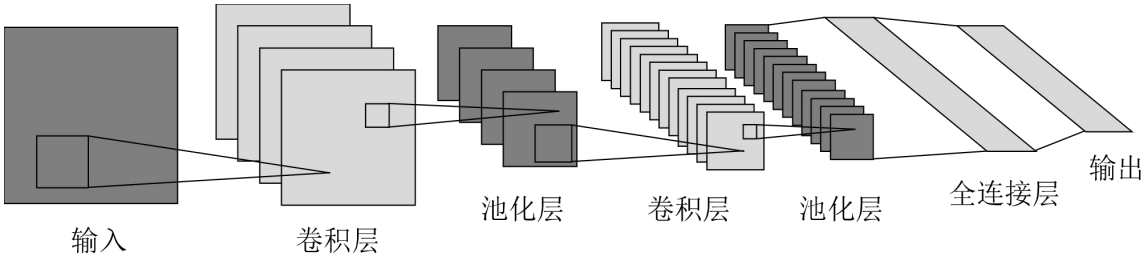


图 2.1 卷积神经网络

Fig. 2.1 Convolutional neural networks

2.1.2 卷积层

卷积层由一个或多个卷积核构成，是卷积神经网络最主要的网络层，主要作用是对数据进行特征提取，以起到简化数据的目的。在二维图像上进行的卷积运算，其原理如公式 2.1 所示：

$$S(i,j)=(I*K)(i,j)=\sum_{m}\sum_{n}I(m,n)K(i-m,j-n) \tag{2.1}$$

其中 I 表示输入数据，K 表示卷积核，运算符“ * ”表示卷积操作。图 2.2 展示了一个 3×3 大小的卷积核的运算过程。每一个单元格表示一个神经元，单元格中的数字表示像素值。图中输入数据的大小为 5×5，卷积核

的步长为 1，即一次卷积运算结束后，卷积核向后移动一个单元格。卷积核的大小即为感受野，在感受野中对输入数据进行卷积运算，图 2.2 中，阴影部分的单元格右下角的数字表示卷积运算的参数，每次卷积运算的结果作为对应位置的输出。卷积层通过使用卷积核对输入数据进行卷积操作，可以提取到输入数据的主要特征，同时对数据进行过滤，减少神经网络的运算量，进而提高运算效率。

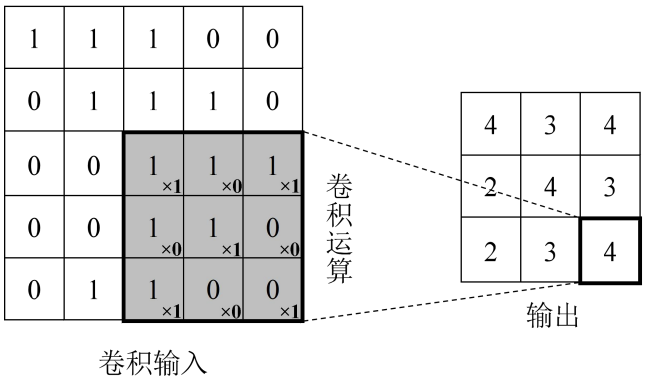


图 2.2 卷积操作

Fig. 2.2 Convolution operation

卷积层的输出称为特征图(Feature map)，对于输入数据来说，每经过一个卷积核的过滤，都会得到一个对应的特征图。一般来说，不同的卷积核可以看作是一种图像模式，卷积核的大小或参数不同，其获取的特征也不同，每一种特征代表着图像某一方面的特性，卷积核越多，其提取到的特征也越丰富，因此，一般都会设置多个卷积核来进行卷积运算，以便获取更丰富的特征。

2.1.3 池化层

池化层主要用来对数据进行降维，从而降低信息的冗余。池化层是模仿人的视觉系统进行的降维，使得网络只关注最有代表性的特征，这种机制称为下采样。神经网络通过池化层的下采样，在对特征进行降维的同时，也能保证网络模型关注的特征不会发生改变，因为网络所关注的是特征是否存在，而不是特征的具体位置。此外池化操作也可以防止一定的过拟合，使网络更容易优化。常见的池化层包括最大池化（Max pooling）、平均池化（Mean pooling）、随机池化（Stochastic pooling）等。最大池化是选取池化层感受野范围内特征的最大值作为池化结果，其它位置的值被舍弃。

平均池化是将感受野中的特征求和后取均值作为池化的结果。在反向传播时，平均池化会将某个值的梯度平均分配给上一层的 n 个值，从而确保池化之后的平均梯度之和不变。随即池化是根据特征区域中各个元素的概率大小来随机选择的，元素值越大，其被选中的概率也越大。池化层不需要训练参数，所以降低了网络的训练复杂度，提升特征提取效率的同时，还可以增强网络的鲁棒性。

2.1.4 全连接层

全连接层的每一个结点都与上一层的所有结点相连，因此叫做全连接层。该层的作用主要是将前面经过卷积层与池化层提取的特征进行整合，并对提取的特征进行分类。使用 ReLU 等激活函数对每个神经元进行激活，可以提升 CNN 的网络性能。一般的全连接层都会引入 dropout 机制[42]，该机制在运算过程中会随机丢弃一些神经元，只对剩余的神经单元进行运算。这种做法是因为尽管之前的卷积操作和池化操作对特征进行了降维，但对于全连接层其获得的特征依然会占用大量的神经元，导致参数量过大，降低了训练效率。随机丢弃一些神经元后，提高训练效率的同时，也可以防止网络出现过拟合，并且可以丢弃神经元之间的一些复杂关系，是网络具有更好的鲁棒性。

2.1.5 几种 CNN 实现网络

(1) LeNet。LeNet 是一种较简单的卷积神经网络，由 LeCun 等人[40]在 1998 年提出，是经典的 CNN 实现模型，可以用来高效地识别手写数字。其网络模型首先有两组卷积层和池化层操作，卷积层中的卷积核大小是 5×5，第一个卷积层会提取 6 个特征图，第二个卷积层会提取 16 个特征图。池化层则使用平均池化，池化单元的大小为 2×2。两组卷积和池化操作后会连接到第三层卷积层，然后连接到全连接层，最终连接到 SoftMax 分类层进行分类。LeNet-5 模型是 CNN 的基础网络模型，它具有结构简单，参数较少，训练速度快的优点。

(2) AlexNet。AlexNet 也是一种结构相对简单的网络模型。是由 Krizhevsky 等人[41]设计提出。AlexNet 使用了 8 层卷积神经网络，前 5 层是卷积神经网络，后 3 层为全连接层。其中，前两层的卷积层都与池化层

相连接，然后连接到后三层的卷积层后再连接到池化层进行池化操作，最后连接到全连接层，并使用分类函数进行分类。AlexNet 是首次使用了 ReLU 激活函数的卷积神经网络，使得提高了网络的训练效率，同时 ReLU 激活函数也能够有效防止网络出现过拟合的情况。此外，AlexNet 使用了层叠池化操作，即池化的大小大于步长，这样可以使得相邻单元之间产生信息交互，且能够保留两者之间的相关联的特性。相比于 LeNet，AlexNet 具有更深的卷积神经网络，且性能比 LeNet 更好，所以也激发了研究人员对于网络深度与性能的关系的研究。

(3) VGGNet。VGGNet 是在 AlexNet 的基础上做出的改进，由牛津大学视觉几何组的 Zisserman 等人提出[43]，VGGNet 网络模型包括了卷积层、池化层和全连接层，除最后的分类层以外，其余所有的全连接层都是采用 ReLU 函数作为激活函数，相比于 LeNet 和 AlexNet，VGGNet 拥有更深的层数，并分别构建了 16 层和 19 层的 CNN。VGGNet 的特点在于，它使用多层连续的较小卷积核来代替较大的卷积核，每一组卷积操作都会使用连续的两层或者三层的卷积层，整个网络的卷积层都采用 3×3 大小的卷积核，并且所有池化层也都使用 2×2 的池化尺寸。VGGNet 的扩展性比较好，其他的研究任务中会直接使用预训练过的 VGGNet 网络模型进行迁移学习，从而完成自己的工作[32]。

(4) GoogleLeNet。GoogleLeNet 是一种更深的网络模型，是由 Google 公司的 Szegedy 等人[44]提出的。GoogleLeNet 使用更深层次的 CNN，具有 22 层的卷积层，并且在网络架构中引入了 Inception 模块，并且对 Inception 模块进行堆叠。虽然该模型的参数量更大，但其效果比之前的卷积网络模型更好。

(5) ResNet。ResNet 是由 He 等人[26]提出的。该模型主要是为了解决深度神经网络出现的退化问题，即随着网络层数叠加到更多层，神经网络的性能反而降低了。ResNet 通过调整网络结构，使其能够更容易的实现优化来解决退化问题。ResNet 将堆叠的一些 layer 作为一个 block，每个 block 可以看作一个函数，并且每个 block 在进行潜在特征学习的同时，也进行了残差学习，使得更深层次的 block 可以学到前面的特征，这样不会使网络的性能出现下降。

2.2 特征融合网络

2.2.1 基本方式

在深度学习中，经常需要进行特征融合，对具有不同维度的特征进行融合，实际上是对输入数据进行信息叠加，使得信息更丰富，从而提升模型性能。特征融合最基本的方式主要有两种，分别是 Add 和 Concat。

对于 Add 融合，可以理解为两个特征直接相加，但只有当待融合的特征具有相同分布，或特征属于同一类时，直接相加才有可能提升模型性能，否则可能导致模型失效。假设两个待融合的特征分别为 X 和 Y，则 Add 融合可记为：

$$X_1' = X + Y \tag{2.2}$$

对于 Concat 融合，是指将两个及以上的特征图在 Channel 或 Num 维度上进行拼接。通道数的合并，意味着描述图像本身的特征增加了，而每一特征下的信息是没有增加的。这种融合方式多用于利用不同尺度特征图的语义信息，以增加通道数的方式实现较好的性能。假设两个待融合的特征分别为 X 和 Y，则 Concat 融合可记为：

$$X_2' = [X:Y] \tag{2.3}$$

假设使用全连接层对上述两种融合后的特征作后续变换，设参数为 W_1，对于 X_1'，则有：

$$Z_1 = W_1 \times X_1' = W_1 \times (X + Y) \tag{2.4}$$

其中 X 与 Y 的维度为 n，设输出维度为 m，则 $W_1 \in \mathbb{R}^{m \times n}$。

对于 X_2' 进行全连接层的线性变换时，设参数为 W_2，则有：

$$Z_2 = W_2 \times X_2' = W_2 \times [X:Y] \tag{2.5}$$

设 X 维度为 n_1，Y 维度为 n_2,同样设输出为 m 维，则 $W_2 \in \mathbb{R}^{m \times (n_1 + n_2)}$。在深度学习中，隐藏神经元数量一般逐步减少，则有 $n_1, n_2 \geq n$，因此 W_2 的参数比 W_1 更多。由于模型的参数越多，模型越复杂，模型的拟合能力越好，所以使用 Concat 融合效果比 Add 融合效果更好。但同时由于 Concat 融合的特征维度更大，其运算量也会增加，所以其训练效率也会受到影响。因此在实际任务中，应根据具体的任务和网络结构来确定使用何种融合方式，并且

通过实验来验证。

2.2.2 特征金字塔网络

特征金字塔网络（FPN，Feature Pyramid Networks）是由 He 等人[45]于 2017 年提出的一种网络，主要用于解决物体检测中的多尺度问题。卷积神经网络随着深度越来越深，其语义信息也越来越丰富，但得到的特征图越来越小，分辨率也越来越低。FPN 网络通过简单的网络连接改变，在基本不增加原有模型计算量的情况下，大幅度提升了小物体检测的性能。

FPN 网络的算法结构如图 2.3 所示。主要包括两个线路，一个自底向上的线路和一个自顶向下的线路，同时在两个线路的每一层都会进行横向连接，图中放大区域即为横向连接，此处使用 Add 方式进行连接，其中 1×1 的卷积核的作用在于减少特征图的个数，且不改变特征图的大小。

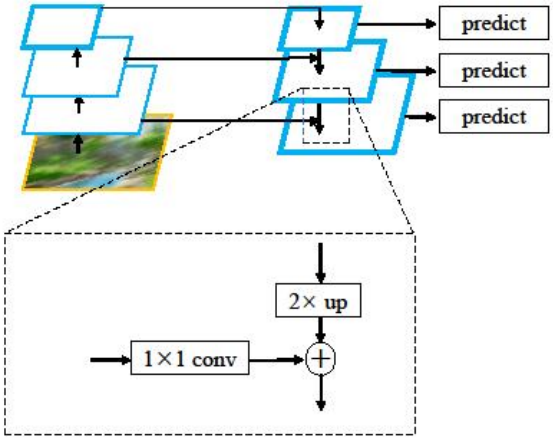


图 2.3 FPN 算法结构示意图

Fig 2.3 Structure diagram of FPN algorithm

在自底向上的过程中，本质就是卷积神经网络前向传播的过程，卷积神经网络中按照特征图的大小划分为不同的 Stage，每个 Stage 的大小比例相差为 2。在自顶向下的过程中，通过上采样的方式，将该层的特征图放大到上一层特征图的大小，这样做是因为可以利用顶层较强的语义特征进行分类，同时也利用了底层的高分辨率信息用于定位。此处上采样使用最近邻插值法，如图 2.4 所示。使用类似于残差网络的连接方式，将上一层经过上采样的特征和当前层的特征进行融合，然后将每一层融合后的特征经过 3×3 的卷积核后进行预测。FPN 网络能够将高层的语义特征与底层的精确定位能力相结合，大大提高了目标检测的准确率。

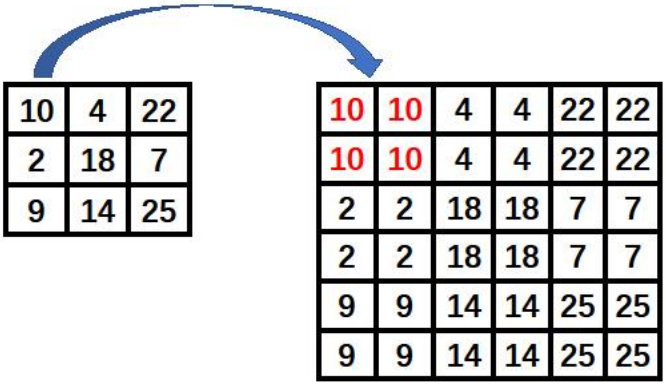


图 2.4 最近邻插值法示意图

Fig 2.4 Schematic diagram of nearest neighbor interpolation method

2.2.3 共享和特定特征迁移网络

共享和特定模态迁移网络（SSTN，Shared and Specific Transfer Network）是 Lu 等人[31]提出的一种用于对跨模态图像特征进行融合的网络，其做法借鉴了 GCN 网络[35]的思想，平滑了各模态的特征，从而进一步挖掘特征之间的联系。该网络主要用于多模态行人重识别任务中对来自不同模态图像的特征进行融合。其网络结构如图 2.5 所示。

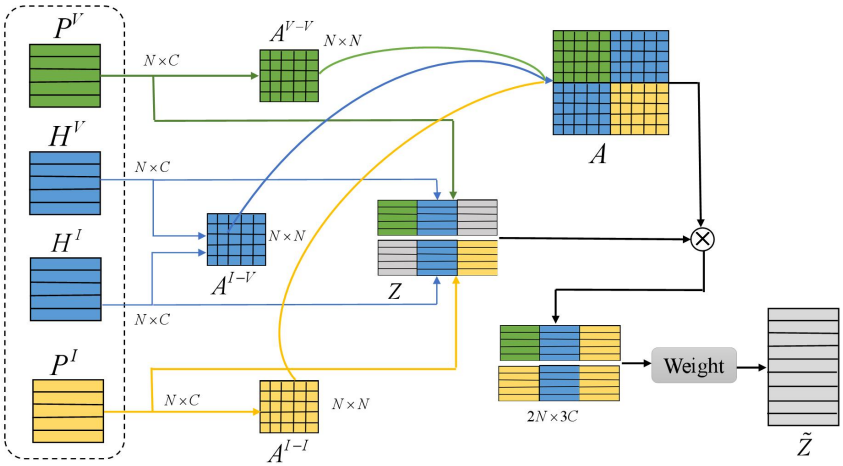


图 2.5 SSTN 网络结构

Fig 2.5 SSTN network structure

其中 P^R、P^I 分别表示来自可将该图像和红外图像的特定特征，H^R、H^I 分别表示来自可将该图像和红外图像的共享特征。使用两种特征来生成亲和矩阵，即：

$$A_{i,j}^{m,m} = d(P_i^m, P_j^m),\ \ A_{i,j}^{m,m'} = d(H_i^m, H_j^{m'}) \tag{2.6}$$

其中 $A_{i,j}^{m,m}$ 表示第 i 个样本和第 j 个样本之间的模态内亲和力，两者属于同一模态，$A_{i,j}^{m,m'}$ 表示模态间亲和力，两者不属于同一模态。$d(a,b)$ 表示归一化的欧几里得距离，如公式 2.7 所示。

$$d(a,b) = 1 - 0.5 \cdot \left\| \frac{a}{\|a\|} - \frac{b}{\|b\|} \right\| \tag{2.7}$$

由此可得最终的亲和矩阵 A，如公式 2.8 所示。

$$A = \begin{pmatrix} \tau(A^{R,R}, k) & \tau(A^{R,I}, k) \\ \tau(A^{I,R}, k) & \tau(A^{I,I}, k) \end{pmatrix} \tag{2.8}$$

其中 $\tau(\bullet, k)$ 为近邻选择函数，其作用是将矩阵的前 k 行保留，并将其它值设为 0。

使用公式 $d_{ii} = \sum_i A_{ij}$ 获取亲和矩阵 A 的对角矩阵 D，然后对拼接的特征矩阵 Z 使用近邻结构（$D^{-\frac{1}{2}} A D^{-\frac{1}{2}} Z$）进行传播，并使用非线性变换进行融合。特征融合后，传播的特征将包括两种模态的共享特征和特定特征。传播特征 Z 的计算如公式 2.10 所示。

$$d_{ii} = \sum_i A_{ij} \tag{2.9}$$

$$\tilde{Z} = \sigma(D^{-\frac{1}{2}} A D^{-\frac{1}{2}} Z W) \tag{2.10}$$

其中 σ 为激活函数，此处使用 ReLU 函数，W 是该网络的可学习参数。最后将所得的传播特征输入到特征学习网络中，以优化整个学习过程。获取的特征记为 T，将 T 作为损失函数的输入来对网络进行训练。

2.3 深度学习中常用技术

前两节介绍了深度学习中的CNN模型和特征融合网络，在深度学习中除了深度神经网络的结构部分，还会涉及到许多非结构的部分，以及一些训练中常用的技术，例如数据增强、激活函数、目标函数优化算法、损失函数等。

2.3.1 常用数据增强方法

在深度学习中，为了提高深度神经网络模型的泛化能力，在数据量有限

的前提下，需要对已有数据进行数据增强，通过对数据的颜色、形状、大小等特征的改变来提高训练数据的多样性，以此来提高网络的泛化能力。根据数据的增强方式主要分为单样本数据增强、多样本数据增强和无监督数据增强。

（1）单样本数据增强。单样本数据增强在一次操作中是只针对一个样本进行的，以图像样本为例，单样本数据增强的方式包括几何变换类、颜色变化类等。其中几何变换类包括对单个图像样本进行翻转、裁剪、变形、旋转、缩放和扩大等。颜色变化类包括随机添加噪声、颜色变换、模糊、填充等。在网络训练中经常会使用几何变换类的数据增强方法将输入图片修改为统一大小的图片作为输入。颜色变换类的增强方法会改变数据本身的内容，因此可以有效提高网络模型的泛化能力。

（2）多样本数据增强。多样本数据增强是利用多个样本来产生一个新的样本的方式。主要包括SMOTE（Synthetic Minority Over-sampling Technique）方法[49]、SamplePairing方法[50]和Mixup方法[51]。SMOTE是专门针对样本空间内的小样本进行数据增强的，该方法基于插值的方式，通过人工合成新样本来解决小样本的不平衡问题。SamplePairing方法是针对于图像样本的一种增强方式，在对两张图片进行基础的几何或颜色变换后，将两张图片的像素值相加并取均值，将该结果作为新的样本，新样本的标签可以为两个原样本标签的任意一个。Mixup方法是一种基于邻域风险最小化原则的数据增强方法，该方法在样本空间中随机选择两个样本，并分别取这两个样本中的一部分拼接成新样本。以上三种方式都是将离散样本连续化来拟合新样本。

（3）无监督数据增强。无监督的数据增强方法是通过模型的学习来生成数据的方式，主要包括GAN（generative adversarial networks）[34]和AutoAugment[52]。其中GAN通过学习数据的分布来生成新的样本，GAN包含了生成网络和对抗网络，生成网络用于生成新的数据，对抗网络用于判断生成的数据是否真实。AutoAugment方法是通过模型学习出适合当前任务的数据增强方法，该方法基本思路是使用增强学习从数据本身寻找最佳的数据变换策略。

2.3.2 常用激活函数

（1）Sigmoid 函数。Sigmoid 函数是一个拥有 S 型曲线的函数，其表达式为公式 2.11，该函数将函数值非线性的映射到(0, 1)区间，其对于绝对值较大的值抑制效果比较强。

$$S(x)=\frac{1}{1+e^{-x}} \tag{2.11}$$

此外，Sigmoid 函数的导数可以比较方便的用自身来表达，如公式 2.12 所示。

$$S'(x)=\frac{e^{-x}}{(1+e^{-x})^2}=S(x)(1-S(x)) \tag{2.12}$$

（2）tanh 函数。tanh 函数表达式为公式 2.13，将函数的值非线性的映射到(-1, 1)的区间上。

$$\tanh(x)=\frac{e^{x}-e^{-x}}{e^{x}+e^{-x}} \tag{2.13}$$

（3）ReLU 函数。ReLU 函数是应用比较广泛的激活函数，其表达式如 2.14 所示，该函数是一个分段函数，有公式可知，ReLU 函数仅对值小于 0 的部分进行了单侧抑制。

$$\mathrm{ReLU}(x)=\begin{cases}x & x>0\\ 0 & x\le 0\end{cases} \tag{2.14}$$

（4）LeakyReLU 函数。LeakyReLU 函数是在 ReLU 函数基础上改进而来，解决了 ReLU 因为抑制造成的有些神经元可能永远不会激活的问题。相比于 ReLU 函数，LeakyReLU 函数不再将负值设定为 0，而是给了一个较小的斜率，这样在进行求导时可以获得一个比较小的梯度。LeakyReLU 函数如公式 2.15 所示，其中$0<\alpha<1$。

$$\mathrm{LeakyReLU}(x)=\begin{cases}x & x>0\\ \alpha x & x\le 0\end{cases} \tag{2.15}$$

2.3.3 常用目标函数优化算法

（1）梯度下降法与随机梯度下降法。梯度下降法(Gradient descent

algorithm, GD)[48]是在求最优化问题中的最基本算法之一，在机器学习中得到了广泛的应用。在进行梯度下降法时，首先需要确定的是当前位置下损失函数的梯度，然后通过梯度和步长确定下一步需要下降的距离，并完成一次迭代，重复上述步骤进行迭代，直到函数收敛为止。由于传统的梯度下降法更新参数时需要通过所有的样本来计算更新梯度，随着如今数据量的增长以及深度学习对大量训练样本的需求，传统的梯度下降法的计算效率显然太低，已经不能满足当今机器学习对于函数优化效率的需要，所以随机梯度下降法(Stochastic gradient descent algorithm, SGD)[46]被提出。随机梯度下降法是基于梯度下降法进行的改进，与梯度下降法相比，随机梯度下降法每次只从训练集中随机选取某个样本进行梯度计算和更新迭代，该算法的计算量小，而且收敛速度更快等，所以可以应用于大规模的数据集中，并且适用于训练样本不断更新的场景。

（2）Adam 优化算法(Adaptive moment estimation)[47]。在使用梯度下降法进行迭代时会存在许多问题。首先，梯度下降法可能会由于选择了不合适的起始梯度方向，对后续的迭代造成的波动会比较大，使得收敛速度比较缓慢。另外梯度下降法是根据梯度方向进行更新的，梯度为 0 或者更新距离较小时则会停止更新，鞍点梯度为 0 但并不是最优解，所以还存在容易陷入鞍点的问题。由于以上这些问题的存在，许多优化算法被提出，在加入动量的随机梯度下降法中，使用动量参数 α 和之前梯度的累计对优化方向进行了调整，而不是单一的使用当前的梯度进行调整，这在一定程度增加了稳定性。因此越来越多的基于动量的优化算法被提出，Adam 优化算法就结合了基于动量的优化算法的优势，该算法可以为不同的参数设计出独立的自适应性学习率。Adam 优化算法对于超参数选择更容易，并且计算简单，计算效率更高，在进行大规模数据量的运算时具有很大的优势，是现今深度学习领域中应用十分广泛的优化算法。

2.3.4 常用损失函数

（1）交叉熵损失函数。交叉熵损失函数常用于分类模型，包括二分类任务和多分类任务，其表达式如 2.16 所示。

$$L(Y, f(x)) = -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k} Y_{ij}\log(f(x)_{ij}) \tag{2.16}$$

（3）平方损失。平方损失对两者距离的平方求和，常用于回归任务。

$$L(Y,f(x))=\sum_{i=1}^{n}(Y_i-f(x)_i)^2 \tag{2.17}$$

（4）L1 损失。L1 损失也就平均绝对误差（MAE），是一种回归损失函数，它表示预测值与真实值之间距离的均值，对于难样本或异常值更具有鲁棒性。其表达式如下所示。

$$L=\frac{\sum_{i}^{n}\left|y_i-y_i^P\right|}{n} \tag{2.18}$$

（5）L2 损失。L2 损失也叫均方误差损失（MSE），是一种解决回归问题的损失函数，在进行梯度更新时，收敛方向更准确，所以收敛速度更快。其表达式如下所示。

$$L=\frac{\sum_{i=1}^{n}(y_i-y_i^P)^2}{n} \tag{2.19}$$

（6）三元组损失。常用于行人重识别任务中。使用 a、p、n 分别表示基准样本、正样本和负样本，则三元组损失可使用公式 2.20 表示。

$$L=\max\{d(a,p)-d(a,n)+m\arg in,0\} \tag{2.20}$$

使用三元组损失函数时，常对样本进行难样本挖掘，以此来提高训练效率，降低训练成本，如 2.21 所示。

$$L=\max\{\max_{p\in A}d(a,p)-\min_{n\in B}d(a,n)+m\arg in,0\} \tag{2.21}$$

2.4 本章小节

本章围绕行人重识别领域涉及到的主要相关技术进行了介绍。首先介绍了深度学习中的卷积神经网络，分别介绍了其组成结构和常见的几种 CNN 网络模型。之后介绍了深度学习中经常使用的一些特征融合算法，包括特征融合的基本方式和两种典型的特征融合网络，同时特征融合网络也是行人重识别任务中比较重要的模块。最后介绍了深度学习中常用的一些技术，包括常用的数据增强方法、常用的激活函数、优化算法以及损失函数。这些技术也是多模态行人重识别技术的基础，本文提出的优化策略也是基于这些基础技术的启发而实现的。

第 3 章 基于行人姿态不对齐的行人重识别数据增强方法

3.1 问题提出

行人重识别工作的现实意义在于从监控图像中找到某个特定身份的人物，在实际应用中，输入神经网络的数据都来源于真实场景下的数据。无论是在科研领域还是在工程领域，大部分神经网络都是依靠公共数据集来训练的。虽然公共数据集可以模拟大部分真实情况下的场景，并且在实际使用中也能取得不错的效果，但相比于真实场景，公共数据集中仍然存在与实际场景不一致的情况。对于多模态行人重识别领域，主要使用两种数据集，分别是 Wang 等人[32]提出的 SYSU-MM01 数据集和 Chen 等人[33]提出的 RegDB 数据集，这两种数据集模拟了真实场景下监控图像中的行人图像，但仍然存在与真实数据不匹配的情况，其中最多的是姿态不对齐问题。在实际场景中，不同的摄像头由于拍摄的角度和距离不同，导致得到的图片与真实的行人大小比例不符，且图片上半部分背景图像占比较大，成为数据集中的难样本，如图 3.1 所示，由于真实场景中会存在大量图像比例与人体比例不一致的现象，而数据集中的图像比例大多都很好的符合了人体比例，且数据集中没有足够的该类图片对网络进行训练，因此使用公共数据集训练的网络在输入真实图像时，在遇到姿态不对齐的数据无法准确的进行识别，会使网络更多的专注于正常比例的数据，进而降低了重识别的准确率。

针对以上的问题，本文提出一种数据增强方法，对简单样本的行人图像进行裁剪和填充，使之成为姿态不对齐的难样本图像，将生成的数据与原数据一起作为输入数据来对网络进行训练。通过对数据进行预处理，使得输入网络的数据能够更好的与实际场景相符，在扩大训练数据的同时，也使得神经网络有更多的难样本数据来训练网络，减小由于行人姿态不对齐所带来的识别准确率降低，进而使得神经网络具有更好的鲁棒性。

图 3.1 姿态不对齐的难样本
Fig 3.1 Difficult sample of attitude misalignment

3.2 相关数据集

当前在跨模态行人重识别领域主要有两个数据集，分别是 SYSU-MM01 数据集[27]和 RegDB 数据集[33]。

3.2.1 SYSU-MM01 数据集

SYSU-MM01 包含由 6 个摄像机拍摄的行人图像，其中 2 个红外摄像机（cam3 和 cam6），4 个可见光摄像机（cam1，cam 2，cam 4，cam 5），数据集信息如表 2.1 所示。该数据集共有 287628 张 RGB 图像和 15792 张红外图像，同时包含了 491 个行人身份，其中训练集包含了 296 个身份，验证集包含 99 个身份，测试集包含 96 个身份。此外，SYSU-MM01 数据集根据拍摄地点的不同设计了两种场景，分别为户内场景和户外场景，户内场景下的 RGB 图像由 cam1 和 cam 2 拍摄，红外图像由 cam3 拍摄。而户外场景下的 RGB 图像由 cam 4 和 cam 5 拍摄，红外图像由 cam6 拍摄。户外场景对于图像识别的挑战性更大。RegDB 数据集包含了 412 个行人身份，每个身份拥有 10 张可见光图像和 10 张红外图像，在这 412 个身份中，有 156 个身份是从正面拍摄，其余 256 个人是从背面拍摄。该数据集图像较小 ，且同一身份的图像姿态变化也较小，因此一定程度上降低了训练难度。示例图像如图 2.1 所示。

表 3.1 SYSU-MM01 数据集主要信息

Table 3.1 Main information of SYSU-MM01 dataset

Cam	location	(in/out)door	lighting	ID	RGB	IR
1	room1	indoor	bright	259	400+	-
2	room2	indoor	bright	259	400+	-
3	room2	indoor	dark	486	-	20
4	gate	outdoor	bright	493	20	-
5	garden	outdoor	bright	502	20	-
6	passage	outdoor	dark	299	-	20

SYSU-MM01 数据集对于在网络训练中的使用也做了一些规定。在测试阶段，将所有 RGB 图像用于图库集（Gallery set），红外图像用于探测集（Probe set）。同时设置了两种模式：全搜索模式（All-search model）和室内搜索模式（Indoor-search model）。对于全搜索模式，RGB 相机 1、2、4 和 5 的图像用于图库集（Gallery set），红外相机 3 和 6 的图像用于探测集（Probe set）。对于室内搜索模式，RGB 摄像头 1 和 2 的图像用于图库集，红外相机 3 和 6 的图像用于探测集。

对于以上两种模式，作者根据其是否使用位于同一位置的摄像头拍摄进行了单镜头和多镜头的设置。例如，对于 RGB 相机下的每个身份，我们随机选择 1 张或 10 张的身份图像来形成单镜头或多镜头的画廊集，对于探测集，则使用全部图像。对于行人重识别任务，给定一个探测图像，通过计算探测图像和图库图像之间的相似性来进行匹配，这种匹配是在不同位置的摄像机之间进行，在 SYSU-MM01 数据集中，由于相机 2 和相机 3 位于同一位置，所以，对相机 3 的探测图像就会跳过了相机 2 的图库图像。

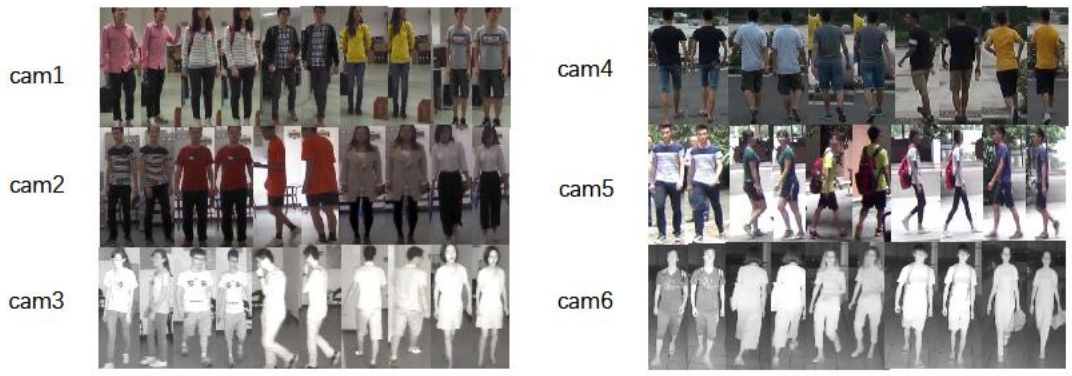


图 3.2 SYSU-MM01 数据集中 RGB 图像与红外图像对比

Fig 3.2 Comparison between RGB images and infrared images in SYSU-MM01 dataset

3.2.2 RegDB 数据集

RegDB 数据集是由 Chen 等人[33]首次提出的一种多模态行人数据集，但它并不是为了解决多模态下的行人重识别问题而提出的，而是为了解决单模态下的行人重识别提出的。由于热力图像中人体与背景的差异特征明显，所以更能凸显人体的姿态信息，对于人体姿态的特征提取来说相对更容易。作者使用热力图像配合 RGB 图像进行训练，将热力图像中的姿态特征与 RGB 图像的特征进行融合，由于热力图像中身体与背景的区别性大于可见光图像，更容易检测行人区域，同时包含了身体姿态的信息，但不包含身体颜色、纹理等信息，在识别时对衣服和环境颜色的变化具有鲁棒性，因此在单模态行人重识别任务中结合热力图像进行识别，其效果有了很好的提升。

RegDB 数据集中的图像是使用普通摄像机和红外摄像机在同一位置对行人同时拍摄所得的，数据集一共有 412 个身份，对于每个人，捕获了 10 个可见光图像和相应的 10 个热图像，因此，数据库包含 4120 个可见光图像和 4120 个对应的热图像，示例图像如图 2.2 所示。由于这些图像是在人们移动时拍摄的，所以每个人的 10 张图像在身体姿势、捕捉距离和光照条件上都存在差异。但是同一个人的 10 幅图像中，相机的天气状况、视角和拍摄视角都是相同的。每个身份的 RGB 图像和热力图像的姿态都是一一对应的，并且同一个身份在姿态上变化很小，这也使得在该数据集上进行跨模态识别时的难度降低。

由于该数据集最初是用来解决单模态行人重识别问题的，所以作者并未给出在跨模态行人重识别任务中的训练方法，在学术界，以下方法是应用比较多的方法：

（1）在训练时随机选择 216 个人物身份 ID，其余的 216 个 ID 用于测试。

（2）使用 RGB 图像作为图库集，热力图像作为探测集进行训练并评估；

（3）使用热力图像作为图库集，RGB 图像作为探测集进行训练并评估；

（4）重复以上三个步骤，随机划分 10 次后，取平均值作为最终结果。

在以上方法中，（2）和（3）分别对应 Thermal to Visible 和 Visible to Thermal 两种模式。

图 3.3 RegDB 数据集 RGB 图像与热力图像对比

Fig 3.3 Comparison between RGB images and thermal images of RegDB dataset

3.3 基于行人姿态不对齐的数据增强方法

在一般的计算机视觉任务中，为了提高模型的泛化能力以及鲁棒性，会对图像进行数据增强，例如裁剪、旋转、填充等。基于数据增强的基本思想，并结合跨模态行人重识别任务中与实际场景的差距，本文提出了一种数据增强方法，增强后的数据可以有效模拟实际场景中行人姿态不对齐的图像，以此来提高重识别网络的泛化能力和鲁棒性。

本文选择每个行人身份所对应图片数量的四分之一作为预处理的数据。首先将数据集中的图像大小统一调整为 144×288，然后将训练数据的大小调整为 108×216，即长和宽缩小四分之一，再将图像左右两侧各填充 18 像素，下方填充 72 像素，使得图像大小统一为 144×288。并随机水平翻转，最后将调整后的图像与原数据一起作为训练数据，数据处理前与处理后的图像对比如图 3.2 所示。通过压缩和填充，得到的图像很好的模拟了行人姿态不对齐的场景。

在实际应用中，行人重识别的数据来源于行人识别阶段，即通过行人识别算法对监控图像或视频进行识别，将识别出的人体图像截取出来作为行人重识别阶段的输入数据。由于经过行人识别算法截取的图像大小有限，同时监控图像的像素大小和画面质量有限，所以获得的图像数据信息量较小。基于以上因素，本文只对图像进行了收缩与填充的操作，并没有使用

裁剪和遮挡等数据增强的手段，以此来保证经过处理后的数据能够最大程度的保留原来数据的信息。

图 3.2 处理前后图像对比

Fig 3.2 Image comparison before and after processing

3.4 实验及分析

本小节主要介绍了对基于行人姿态不对齐的数据增强方法进行的相关实验。该部分内容主要包括实验相关模型、实验相关设置、实验评价指标、实验环境以及实验结果与分析几个部分。

3.4.1 实验相关模型

为了证明本文提出的数据增强方法对于行人重识别模型的泛化能力，本文选取了多模态行人重识别领域内的多种结构的神经网络模型进行了实验，包括单流结构、双流结构、非对称全连接层结构。其中，单流结构选取了深度零填充算法的单流网络模型，双流结构选取了模态感知协作网络（MACE）[28]和跨模态特征转移网络（cm-SSFT）[31]，非对称全连接层结构选取了一致性正则化的非对称距离模型（CVDCA）[33]，同时也选取了多模态行人重识别领域中结合了生成对抗网络[34]（GAN, Generative Adversarial Networks）的 cmGAN[35]和 AlignGAN[36]。以上 6 种结构的网络模型是多模态行人重识别领域表现比较出色的模型，并且可以作为该领域

中各种结构类型的代表性网络模型，使用以上模型作为实验模型，可以很好的证明基于行人姿态不对齐的数据增强方法对于提高模型泛化能力的有效性。

3.4.2 实验相关设置

实验数据集使用 SYSU-MM01 数据集和 RegDB 数据集。对于 SYSU-MM01 数据集，分别使用了室内搜索模式和全搜索模式进行了实验。对于 RegDB 数据集，使用 3.2.2 介绍的方法进行实验。实验之前，将输入图片的大小设置为 288×144，并按照 3.3 中的方法进行数据增强。在每次训练中随机选取 P=8 个身份标签，然后在数据集中随机选取对应身份的 K=4 个可见光图像及 K=4 个红外图像，即每个批次训练包含 32 张可见光图像和 32 张红外图像，总的训练批次大小为 64。训练迭代次数为 60，学习率在前 10 次迭代中由 0.01 递增到 0.1，在第 10 到第 30 次迭代中保持为 0.1，30 次以后为 0.01。

3.4.3 实验评价指标

本文的实验使用累积匹配特征（CMC）和平均精度均值（mAP）作为性能评价指标，同时这两个标准也是行人重识别领域主要使用的性能评价指标。

3.4.3.1 累积匹配特征

CMC，全称累计匹配特征（Cumulative Matching Characteristics），主要用于行人重识别领域中的性能评估，同时也算模式识别系统中的重要评价指标。行人重识别任务中，输入一组 Gallery 图像，输出的结果会将 Gallery 图像根据与 Query 图像的距离由小到大排序，当 Query 集中只有一个 ID 实例时，CMC 的计算公式如下：

$$A_{cck}=\begin{cases}1, & \text{排名前}k\text{的gallery样本包含query标识}\\ 0, & \text{其它}\end{cases} \tag{3.1}$$

式 3.1 是一个阶跃函数，大部分场景下 Query 集合中不止一个 ID 实例，因此会将每个 Query 的 A_{cck} 相加，再除以 Query 中 ID 的总数，得到平均 A_{cck} 曲线。

由于 CMC 曲线是一个单位阶跃函数，因此必定是一个单调递增的曲线。曲线上某个点如(R5, 0.96)就表示正确结果在前返回所有结果中排名前 5 的准确率能达到 96%。在实际论文中，一般只取 CMC 曲线上的某几个点来进行对比，比如经常使用的 rank-1、rank-5、rank-10 等，就分别是 CMC 曲线上 k=1、5、10 时的值。

在模式识别和行人重识别任务中，rank-k 是指搜索结果中置信度最高的前 k 张图中有正确结果的概率。以行人重识别任务为例，根据与 Query 中标签图像的相似度来对 Gallery 集合的图像进行由小到大的排序，设 Query 中有 n 个标签需要识别，则 rank-k 的计算过程如下：

（1）计算 Gallery 数据集中每个图像与 Query 中对应标签的相似度。

（2）根据相似度对 Gallery 集合进行降序排序。

（3）判断与 Query 的标签相同的样本是否存在前 k 个位置中，如果存在，则为 True，反之则为 False。

（4）统计步骤 3 中 True 的个数，记为 m，则 rank-k=m/n。

3.4.3.2 平均精度均值

mAP，全称平均精度均值（mean Average Precision）。使用 P 表示准确率，R 表示召回率，则 mAP 的规范表示为：

$$\mathrm{mAP}=\int_0^1 P(R)dR \tag{3.2}$$

由上式可知，mAP 表示的是 P-R 曲线与坐标轴之间形成区域的面积，直观的表述为，一个系统的性能较好，其 P-R 曲线是尽可能的向上凸出的，即准确率更高，召回率更低。

在行人重识别任务中，mAP 是通过计算准确率来得到的。设 query 集合中有 n 张图像，Query 图像对应的 Gallery 集合分别有 $m_1,m_2,\ldots,m_n$ 张正样本，则 mAP 计算过程如下：

（1）依次选取 Query 集合中的样本作为查询图像 q。

（2）在重新排序后的 Gallery 集合中，找到其中是查询图像 q 的正样本的所有位置，计算每个位置下的准确率 AP_i。

（3）将每个位置下的准确率 AP_i 相加并取均值，作为查询图像 q 的准确率 AP_q。

（4）重复步骤（1）~（3），直至遍历完所有 Query 图像。

（5）将所有 AP_q 相加并取均值，即求得 mAP。

3.4.4 实验环境

本章实验的实验软硬件环境如表 3.1 所示。

表 3.1 软硬件环境

Table 3.1 Software and hardware environment

硬件	CPU	Inter Core i7-8700 CPU @ 3.2GHz
	GPU	NVIDIA RTX 2080Ti 11GB
	内存	16G
	硬盘	1TB
软件	操作系统	64 位 Ubuntu 16.04 系统
	开发语言	Python 3.6
	开发框架	Pytorch 1.0.1
	开发工具	PyCharm 2019.03

3.4.5 实验结果与分析

本节给出了使用 3.4.1 中的网络模型分别在数据集进行数据增强前后的实验结果数据。实验模型包括单流结构的深度零填充模型（deep zero-padding）、双流结构的模态感知协作网络（MACE）和跨模态特征转移网络（cm-SSFT）、非对称全连接层结构中的一致性正则化的非对称距离模型（CVDCA）以及使用了对抗生成网络的 cmGAN 和 AlignGAN。评价标准使用 CMC 中的 rank-k 和 mAP。

本章分别在 SYSU-MM01 数据集的两种模式下进行了实验，表 3.2 为 SYSU-MM01 数据集在 All-Search 模式下数据增强前后各个模型的实验结果，表 3.3 为 Indoor-Search 模式下的实验结果，两种模式中都包括单镜头模式和多镜头模式。

表 3.2 SYSU-MM01 数据集 All-Search 模式下的实验结果

Table 3.2 Experimental results of SYSU-MM01 dataset in all-search mode

Model		All-Search							
		Single-shot				Multi-shot			
		r1	r10	r20	mAP	r1	r10	r20	mAP
Zero-Padding	before	14.8	54.1	71.3	15.9	19.1	61.4	78.4	10.9
	after	11.6	52.8	74.2	13.2	15.7	59.0	64.5	6.6
MACE	before	45.3	79.8	91.1	56.9	51.6	87.3	94.4	50.1
	after	43.6	76.5	86.3	54.1	47.1	83.7	90.3	47.2
cm-SSFT	before	61.6	89.2	93.9	63.2	63.4	91.2	95.7	62.0

	after	58.4	86.7	89.9	61.4	59.6	88.4	92.9	59.6
CVDCA	before	9.3	43.3	60.4	10.8	13.1	52.1	69.5	6.7
	after	7.8	37.2	54.1	8.2	9.4	47.5	64.2	4.4
cmGAN	before	27.0	67.5	80.6	27.8	31.5	72.7	85.0	22.3
	after	25.7	62.5	74.8	23.1	27.6	69.5	81.9	18.8
AlignGAN	before	42.4	85.0	93.7	40.7	51.5	89.4	95.7	62.0
	after	36.6	80.1	89.6	38.2	46.7	84.5	90.4	54.1

表 3.3 SYSU-MM01 数据集 Indoor-Search 模式下的实验结果

Table 3.3 Experimental results of SYSU-MM01 dataset in Indoor-search mode

Model		Indoor-Search							
		Single-shot				Multi-shot			
		r1	r10	r20	mAP	r1	r10	r20	mAP
Zero-Padding	before	20.6	68.4	85.8	26.9	24.4	75.9	91.3	18.6
	after	15.8	64.1	81.5	23.1	20.6	71.9	88.3	14.7
MACE	before	48.3	81.8	91.6	66.2	57.4	93.0	97.5	64.8
	after	46.5	78.7	88.4	63.1	53.9	88.7	94.6	61.4
cm-SSFT	before	70.5	94.9	97.7	72.6	73.0	96.3	99.1	72.4
	after	66.3	90.6	92.8	69.3	69.1	92.5	95.9	69.1
CVDCA	before	14.6	57.9	78.7	20.3	20.1	69.4	85.8	13.0
	after	11.7	54.1	74.8	17.1	16.9	64.2	81.4	9.5
cmGAN	before	31.6	77.2	89.2	42.2	37.0	80.9	92.1	32.8
	after	28.8	72.6	85.7	38.1	34.8	76.9	88.1	28.7
AlignGAN	before	45.9	87.6	94.4	54.3	57.1	92.7	97.4	45.3
	after	42.7	82.1	90.7	50.9	54.0	88.3	93.6	41.5

由上述实验结果可知，不同结构的多模态行人重识别网络在使用数据增强后的 SYSU-MM01 数据集训练后，其 rank-k 和 mAP 均有所下降。可能的原因是：在使用数据增强之前，训练所得的网络对于姿态对齐的正常图像存在一定的过拟合，当数据中出现姿态不对齐的图像时，就会导致其准确率下降。对于在实验中使用数据增强，仅是降低了模型的过拟合程度，并没有提升网络模型的识别能力，因此还需要对行人重识别网络在表征学习和度量学习两个方面进行优化。

其次，由上述结果可知，All-Search 模式下的准确率依然低于 Indoor-Search 模式下的准确率，原因是 Indoor-Search 模式下的图像均来自室内摄像头拍摄，相比于 All-Search 模式下包含的室外图像，其图像背景相对更单一，对于行人重识别网络容易学习到更稳定的特征，而使用数据

增强后，两者图像背景的差别依然存在，所以 Indoor-Search 模式下的准确率更高。

同时由上述两表可知，无论是 All-Search 模式还是 Indoor-Search 模式，单镜头模式的识别准确率总是低于多镜头模式的识别准确率，其原因是：在单镜头模式下，使用的 Query 图像只来自一个摄像头，多镜头模式下的 query 图像来自多个镜头，因此在识别过程中对于 Query 图像提取的特征，多镜头模式下 Query 图像的特征要比单镜头模式下的 Query 图像特征更丰富，所能匹配到的正样本也越多，所以导致多镜头模式的识别准确率高于单镜头模式的识别准确率。

在上述实验中使用的六种网络模型是多模态行人重识别领域中有代表性的几种模型，其中属于双端共享网络结构的 MACE 网络模型和 cm-SSFT 网络模型相对于其它网络模型的识别效果更好，也是本文主要研究的网络模型。由实验结果可知，上述两种模型在使用数据增强进行训练后，识别准确率均有所下降，但其识别效果相较于其它网络模型依然更好，也进一步证明了双端共享网络模型在行人重识别任务中的有效性。

表 3.4 记录了 RegDB 数据集在数据增强前后各模型的实验结果，其中包含了 Visible to Thermal 和 Thermal to Visible 两种实验方法。由于 RegDB 数据集中热力图像成像模糊，相较于 SYSU-MM01 数据集识别率较低，因此只选取了参考价值更大的 rank-1 和 mAP。

表 3.4 RegDB 数据集实验结果

Table 3.4 Experimental results of RegDB dataset

Model		Visible to Thermal		Thermal to Visible	
		r1	mAP	r1	mAP
Zero-Padding	before	17.8	18.9	16.6	17.8
	after	14.6	14.3	13.7	14.9
MACE	before	72.4	69.1	72.1	68.6
	after	68.6	67.1	67.3	65.8
cm-SSFT	before	72.3	72.9	71.0	71.7
	after	69.8	69.9	67.3	68.9
AlignGAN	before	57.9	53.6	56.3	53.4
	after	52.7	49.2	52.0	49.8

由上述实验结果可知，使用对 RegDB 数据集进行数据增强后，训练所得的网络其准确率也有所下降。其原因是数据增强之前训练所得的网络存在过拟合。同时由上表可看出，在使用 RegDB 数据集实验时，Visible to

Thermal 方式下的识别准确率要高于 Thermal to Visible 模式下的识别准确率。分析原因为：Visible to Thermal 模式下是使用可见光图像作为 Query 图像，以红外图像作为 Gallery 集合，网络可以获得包含信息更多的 Query 特征来进行检索。而在 Thermal to Visible 模式下，网络模型以红外图像作为 Query 图像，其本身信息含量较少，因此网络所能获取的信息也少。对于网络来说，使用丰富的特征在较单一的特征集合中检索，要比使用单一的特征在复杂的特征集合中检索更容易。

对于上述两个实验，对数据集进行数据增强后训练所得网络的识别准确率均有所下降，虽然一定程度上降低了模型对于正常图像的过拟合程度，但并没有提高网络模型的识别能力，因此还需要对网络模型进行进一步优化。本文也将分别从表征学习和度量学习的角度出发，在本章的基础上对网络模型优化，从而提高多模态下行人重识别网络的识别能力。

3.5 本章小结

本章主要针对多模态行人重识别任务中行人姿态不对齐的问题对数据进行了数据增强，通过收缩和填充模拟真实场景下行人姿态不对齐的场景。然后使用增强后的数据进行了实验，实验中分别比较了多种类型的多模态行人重识别网络模型在使用数据增强前后的实验数据。实验表明，使用增强后的数据进行网络训练，其 rank-k 和 mAP 均有一定程度的下降，分析原因为通过数据增强来训练网络，一定程度上降低了网络的过拟合，因此其识别效果会有所下降。使用该方法来对数据进行增强，可以提高网络的鲁棒性，从而在应对数据中出现的行人姿态不对齐时拥有更好的泛化能力。

第 4 章 基于非局部注意力机制的多模态行人重识别模型

在基于双端网络的多模态行人重识别任务中，需要对来自两种模态的特征进行融合，并对融合后的特征进行特征提取。在对融合后的特征进行特征提取时，由于存在长距离依赖，一般的特征提取网络很难准确地提取到相应的特征，特征提取效率不高，从而间接影响了整个算法的准确率。本章使用了非局部注意力网络来优化共享特征的提取网络，使之能够有效地捕捉长距离依赖。同时对于融入了非局部注意力机制的网络模型进行了实验，并对实验结果进行了分析。

4.1 问题提出

多模态下的行人重识别任务相对于单模态下的行人重识别任务其准确率更低，主要原因是多模态场景下无法提取更多的共享特征，而在单模态下可以通过一些优秀的特征提取网络进行特征提取，且本身不存在特征融合的问题。因此，想要提高多模态行人重识别任务的准确率，在选取合适的网络结构的同时，还应该尽可能多的提取到两种模态的共享特征。

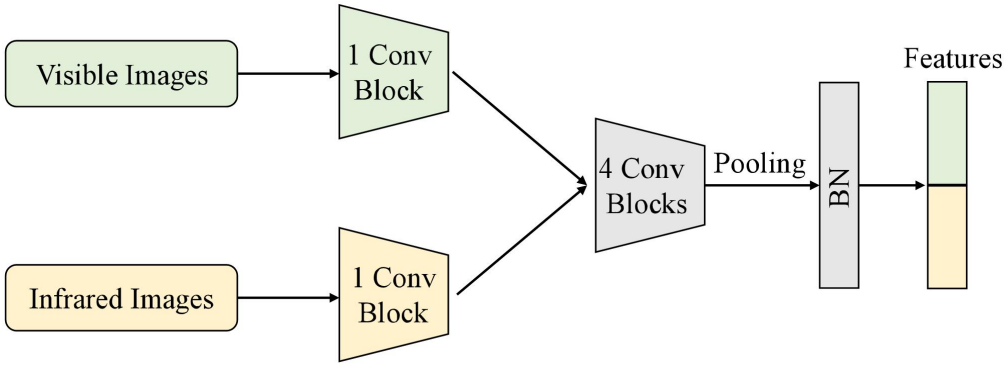


图 4.1 双端网络中的 ResNet-50 特征提取网络

Fig 4.1 ResNet-50 feature extraction network in a two-stream network

基于双端网络的多模态行人重识别模型一般实验 ResNet-50 来进行特征提取。如图 4.1 所示，ResNet-50 拥有五层卷积模块（Conv Block），双端网络将其分为两部分，使用第一层卷积模块作为每个特定模态的特征提

取网络，获得特定特征后，会对两类特征进行拼接，ResNet-50 的后四层卷积模块作为共享网络，将拼接后的特征输入到共享网络中进行共享特征的提取。在进行共享特征提取的过程中，其主要难点在于在拼接后的特征图中，如何捕获大量的长距离依赖。如图 4.2 所示，在拼接后的输入数据中，同样属于人体头部的部位属于相关联的长距离依赖，想要学习到两者之间的内在联系，感受野应该同时包含这两个部分，但在一般的卷积神经网络中，卷积核的大小有限，对于这种空间上的长距离依赖很难一次性捕获，导致最终提取的特征图中无法将这两者之间关联，网络也无法学习到相关联的信息。

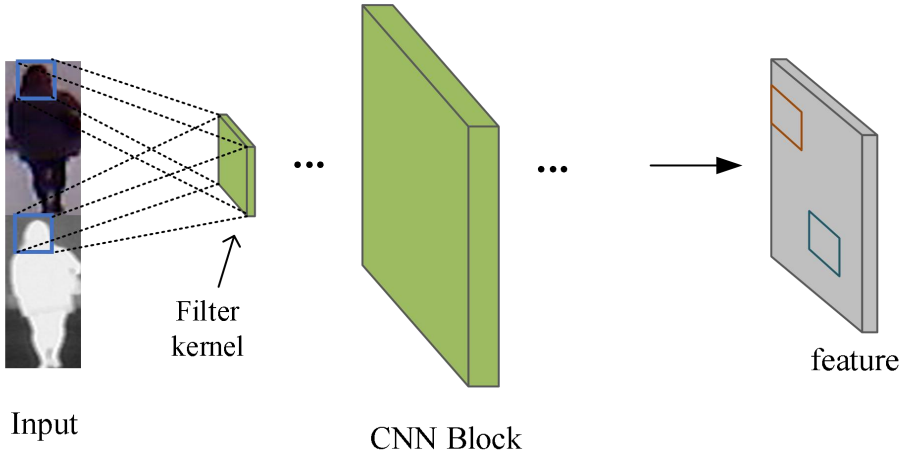


图 4.2 长距离依赖与特征提取

Fig 4.2 Long-distance dependency and feature extraction

对于可见光图像和红外图像，其两者之间的共享特征主要包括人体形态特征及人体部位特征，由于输入到共享特征网络的是拼接后的数据，这就使得合并了两种模态的数据中一些相似的特征在空间上存在一定的距离，在捕获这两种存在依赖的特征时，卷积网络需要更大的感受野才能同时感知到拼接的矩阵中这些特征的存在。而一般的卷积核大小设置的都很小，这导致其感受野只能感知到非常有限的局部区域，想要捕获更多的长距离依赖，只能将感受野扩大到整个矩阵上，使得矩阵中每个点都能参与运算，这样才可以引入全局信息，进而将全局信息传到后面的层。然而将卷积核的大小扩大到整个矩阵是不现实的，在深度学习领域中，非局部注意力机制可以很好地解决这一问题，在扩大感受野并捕获长距离依赖的同时，还可以保证输入前后矩阵的大小保持一致。本章将使用非局部注意力网络对多模态下的双端可共享网络进行改进，以此来提高网络模型的特征提取能力，进而提高行人重识别的准确率。

4.2 非局部注意力网络

非局部注意力网络[36]（Non-local Attention Network）主要用于扩大神经网络的感受野，传统卷积核的感受野大小等于其自身的大小，而一般的神经网络里卷积核大小设置的都很小，这导致其感受野只能感知到非常有限的局部区域，非局部注意力网络将感受野扩大到整个矩阵上，使得矩阵中每个点都能参与运算，这样可以引入全局信息，进而将全局信息传到后面的层，因此非常适合捕获数据中的长距离依赖。使用公式 4.1 可以使得非局部注意力网络的输出在包含全局信息的同时还有和原图一样的大小。

$$y_i = \frac{1}{C(x)} \sum_{\forall j} f(x_i, x_j) g(x_j) \tag{4.1}$$

其中 i 表示输出位置的索引，使用 j 来枚举所有位置。x 表示输入，包括图像、视频、序列号信号和特征等，y 表示与 x 同样大小的输出。二元函数 f 用于计算两者之间的某个标量，如亲和关系等。函数 g 用于计算位置 j 处输入信号的表示，一般使用一个线性函数来表示，可以使 $g(x_i) = W_g * x_j$，其中 W_g 是一个可学习的权重矩阵。$C(x)$主要对结果进行归一化。公式 4.1 中的非局部行为是由于操作中考虑了所有位置，即$\forall j$，而一般的卷积操作只考虑了其卷积核本身大小的位置。

二元函数 f 可以有多种选择，如使用高斯函数，则函数 f 如公式 4.2 所示，正规化因子 $C(x)$如公式 4.3 所示。

$$f(x_i, x_j) = e^{x_i^T x_j} \tag{4.2}$$

$$C(x) = \sum_{\forall j} f(x_i, x_j) \tag{4.3}$$

使用嵌入式形式的点积操作，则函数 f 如公式 4.4 所示，此时为了简化计算，正规化因子 $C(x)$使用 N 表示，代表 x 中位置的数量。

$$f(x_i, x_j) = \theta(x_i)^T \phi(x_j) \tag{4.4}$$

将上述操作进行模块化后，可以直接加入到现有的网络模型中，其模块化模型如图 4.3 所示。非局部注意力网络的模块化操作可使用公式 4.5 表示。

$$z_i = W_z y_i + x_i \tag{4.5}$$

其中 y_i 由公式 4.1 计算所得，与 x_i 相加作为残差连接，使得在加入非局部注意力模块后，不会影响原始网络的初始表现。

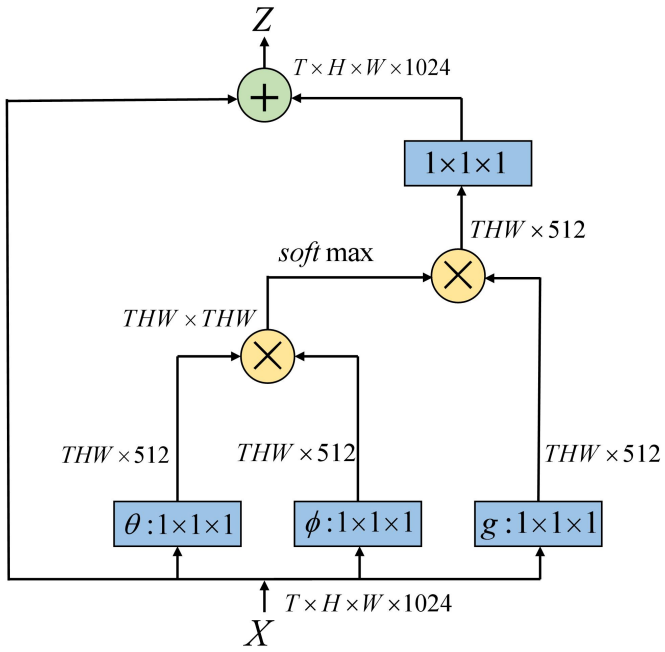


图 4.3 非局部注意力网络结构

Fig 4.3 Structure of nonlocal attention network

4.3 基于非局部注意力机制的 MACE 算法模型

MACE 算法是由 Ye 等人[28]提出的一种模态感知协作算法，该算法是一种双端网络结构的神经网络模型，该算法使用了协同一致性损失及交叉熵损失来训练网络。同时该网络模型使用 ResNet-50 作为特征提取网络，并采用图 4.1 的结构来获取不同模态的特征。本节将非局部注意力网络嵌入到 MACE 网络模型中，以此来提升网络对于共享特征的提取能力。

4.3.1 网络结构

MACE 网络中的共享网络是由 ResNet-50 的后四个卷积模块组成的，本文将非局部注意力网络分别插入到四个卷积模块之后，以此来获得长距离依赖。同时将原有的模态分类器去除，让网络更多的关注于图像的内容而非模态，降低了训练复杂度。改进后的网络模型如图 4.4 所示。

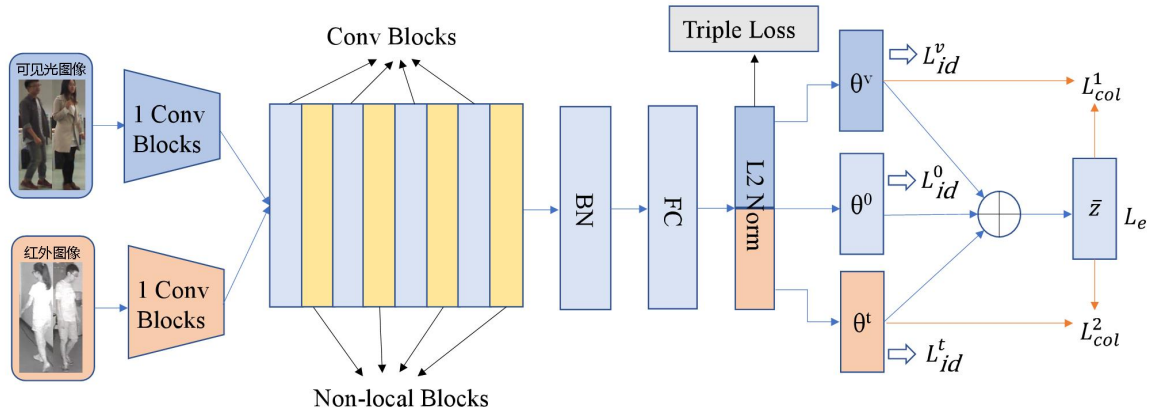


图 4.4 嵌入非局部注意力机制的 MACE 网络结构

Fig 4.4 MACE network architecture embedded with non-local attention block

不同模态的图像经过各自模态的特征提取网络后，将提取到的特征进行拼接，然后输入到嵌入非局部注意力模块的共享网络中，由于非局部注意力模块的加入，使得在拼接的特征图中那些来自不同模态的长距离依赖得以被捕获到。相较于之前的网络，其特征提取能力会有大幅度的提高。经过四层卷积模块和非局部注意力模块的处理，最终获得多模态图像的共享特征。将获得的共享特征进行 Batch Norm 层和全连接层的处理后，输入到 L2 Norm 层，并使用三元组损失进行训练。其次，将特征进行划分后输入到不同模态的分类器中对特征进行分类，分类器使用交叉熵损失来训练，同时为了便于分类器之间的知识转移，使用集成学习损失 L_e 和一致性损失 L_{col} 来训练分类器。

4.3.2 损失函数

如 4.3.1 所述，MACE 算法主要使用四种损失函数来指导网络的训练，分别使用三元组损失指导特征提取网络的训练，使用交叉熵损失、集成学习损失和一致性损失指导分类器的训练。

4.3.2.1 三元组损失函数

三元组损失主要通过使正样本与基准样本之间的距离越来越近，负样本与基准样本之间的距离越来越远来训练网络。多模态下使用三元组损失需要考虑跨模态下的样本差异，由于存在模态差异，跨模态下的正样本与基准样本的距离要比同一模态的正样本与基准样本的距离更大，因此在进

行难样本挖掘时，通常只考虑跨模态下的样本距离。同时考虑到基准样本可能存在不同模态，还采用了双向训练策略[37]来提高性能，它同时考虑了可见光图像到红外图像和红外图像到可见光图像的关系。计算公式如公式 4.6 所示。

$$\begin{aligned} L_{tri} = &\sum_{i=1}^{n}[\rho + \max_{\forall y_j = y_i} D(f_i^v, f_j^t) - \min_{\forall y_i \neq y_k} D(f_i^v, f_k^t)]_+ \\ &+ \sum_{i=1}^{n}[\rho + \max_{\forall y_j = y_i} D(f_i^t, f_j^v) - \min_{\forall y_i \neq y_k} D(f_i^v, f_k^t)]_+ \end{aligned} \tag{4.6}$$

其中 $[\bullet]_+ = \max(\bullet, 0)$，$\rho$为阈值，$n$ 为一个训练批次中输入图像的总个数，f_i^v（f_i^t）表示输入可见光图像（红外图像）后提取的对应模态的特征，$D(\cdot,\cdot)$表示提取的两个样本特征之间的欧式距离的平方[38]。

4.3.2.2 其它损失函数

MACE 算法模型中拥有三个分类器，分别是可见光特征分类器、红外特征分类器和共享特征分类器，使用交叉熵损失函数来作为分类损失函数。使用 $p^0(y_j|x_i^0)$表示输入样本 x_i^0 被识别为 j 的概率，其中上标 0 表示共享特征分类器，y_j 表示对应的标签，其数学表达式如式 4.7 所示。

$$p^0(y_j|x_i^0) = \frac{\exp(z_{i,j}^0)}{\sum_{k=1}^{C}\exp(z_{i,k}^0)}, j = 1, \cdots, C \tag{4.7}$$

式中 $z_{i,j}^0$ 表示样本 x_i^0 经过 θ^0 分类器被认为是 j 的概率。三个分类器的分类损失分别为：

$$\begin{aligned} L_{id}^0 = &-\frac{1}{n}\sum_{i=1}^{n}\log(p^0(y_i|x_i^V)) \\ &-\frac{1}{n}\sum_{i=1}^{n}\log(p^0(y_i|x_i^I)) \end{aligned} \tag{4.8}$$

$$L_{id}^V = -\frac{1}{n}\sum_{i=1}^{n}\log(p^V(y_i|x_i^V)) \tag{4.9}$$

$$L_{id}^I = -\frac{1}{n}\sum_{i=1}^{n}\log(p^I(y_i|x_i^I)) \tag{4.10}$$

上式分别表示共享特征分类器损失、可见光特征分类器损失和红外特征分类器损失，加入权重系数后，则总的分类损失为：

$$L_{id} = L_{id}^{0} + \lambda(L_{id}^{V} + L_{id}^{I}) \tag{4.11}$$

同时使用交叉熵损失指导整体训练，即：

$$L_e = -\frac{1}{n}\sum_{i=1}^{n}\log(p^e(y_i|x_i^V, x_i^I)) \tag{4.12}$$

为了使多个分类器之间协同优化，将不同分类器的输出进行集成，分别使用 $z_i^{0,1}$、$z_i^{0,2}$ 表示样本对 i 经过特征提取网络从可见光图像中获得的共享特征和从红外图像中获得的共享特征，z_i^V 和 z_i^I 表示从对应模态获得的特定特征，取均值后得：

$$z_i^e = \frac{1}{4}(z_i^{0,1} + z_i^{0,2} + z_i^V + z_i^I), i = 1, 2, \cdots, n \tag{4.13}$$

使用知识蒸馏技术进行协作学习，有利于不同分类器之间的知识转移，使用温度参数 T 来平滑不同分类器的概率分布：

$$\tilde{p}^e(y_k|x_i^V, x_i^I) = \frac{\exp(z_{i,j}^e / T)}{\sum_{k=1}^{C}\exp(z_{i,k}^e / T)}, j = 1, \cdots, C \tag{4.14}$$

为了保证特定模态分类器与总分类器之间有相同的分布，引入协同一致性损失，即：

$$\begin{aligned} L_{col} = &\frac{1}{\mathrm{n}}\sum_{i=1}^{n}\sum_{k=1}^{C}\tilde{p}^e(y_k|x_i^V, x_i^I)\log\frac{\tilde{p}^e(y_k|x_i^V, x_i^I)}{\tilde{p}^V(y_k|x_i^V)} \\ &+\frac{1}{\mathrm{n}}\sum_{i=1}^{n}\sum_{k=1}^{C}\tilde{p}^e(y_k|x_i^V, x_i^I)\log\frac{\tilde{p}^e(y_k|x_i^V, x_i^I)}{\tilde{p}^I(y_k|x_i^I)} \end{aligned} \tag{4.15}$$

4.4 基于非局部注意力机制的 cm-SSFT 算法模型

cm-SSFT 算法是 Lu 等人[31]提出的一种使用了特征融合算法 SSTN 的双端行人重识别网络，该网络模型不仅关注各模态图像的共享特征，也关注每个模态的特定特征。本文将非局部注意力网络嵌入到该网络模型中，用以提高共享特征提取网络的特征提取能力。

4.4.1 网络结构

使用非局部注意力网络对模型进行改进后，其网络结构如图 4.5 所示。该网络同样使用双端网络结构，不同于 MACE 网络模型，cm-SSFT 网络在提取共享特征的同时也提取了两种模态下的特定模态特征，并使用 SSTN

算法对三种特征进行融合，SSTN 算法的网络结构如图 2.5 所示。与 4.2 小节一致，本文将共享特征提取网络中的卷积模块嵌入了非局部注意力网络，用以增强共享网络的特征提取能力，捕获长距离依赖。而对于该模型中提取特定模态特征的特征提取网络并没有使用非局部注意力网络进行优化，原因在于相比于共享网络，特定特征网络并不存在长距离依赖，因此不需要非局部注意力网络的参与。

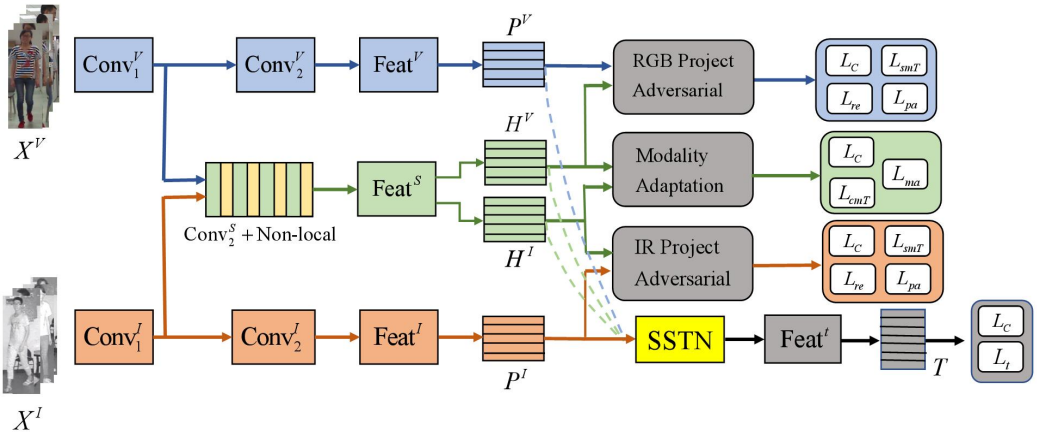


图 4.5 嵌入非局部注意力机制的 cm-SSFT 网络结构

Fig 4.5 cm-SSFT network architecture embedded with non-local attention block

4.4.2 损失函数

在双端网络对图像进行特征提取之后，得到了两类特征，分别是单模态特征 P^R、P^I，跨模态的共享特征 H^R、H^I。为了让这两种特征具有区别性，使用分类损失 L_{id} 和三元组损失 L_{smT}、L_{cmT} 对特征提取网络进行训练，其中 L_{smT} 单模态下的三元组损失，L_{cmT} 表示跨模态下的三元组损失。计算公式如下：

$$L_{id}(H^m) = -\frac{1}{n}\sum_{i=1}^{n}\log(p(y_i^m \mid H_i^m)) \tag{4.16}$$

$$L_{id}(P^m) = -\frac{1}{n}\sum_{i=1}^{n}\log(p(y_i^m \mid \mathrm{P}_i^m)) \tag{4.17}$$

$$\begin{aligned} L_{smT}(P) = &\sum_{i,j,k}\max[\rho_2 + \left\|P_i^R - P_j^R\right\| - \left\|P_i^R - P_k^R\right\|, 0] \\ &+ \sum_{i,j,k}\max[\rho_2 + \left\|P_i^I - P_j^I\right\| - \left\|P_i^I - P_k^I\right\|, 0] \end{aligned} \tag{4.18}$$

$$\begin{aligned} L_{cmT}(H) = &\sum_{i,j,k} \max[\rho_1 + \left\| H_i^R - H_j^I \right\| - \left\| H_i^R - H_k^I \right\|, 0] \\ &+ \sum_{i,j,k} \max[\rho_1 + \left\| H_i^I - H_j^R \right\| - \left\| H_i^I - H_k^R \right\|, 0] \end{aligned} \tag{4.19}$$

其中 $p(y_i^m|\bullet)$ 表示输入图像 X 被分类为 y_i^m 的概率，ρ_1、ρ_2 表示阈值，i、j、k 分别表示基准样本、正样本和负样本的标签。

由于提取到的共享模态特征很容易会包含大量的特定模态特征信息，特定模态特征也并非完全的由特定模态所具备，因此需要对这两类特征做解纠缠，利用模态自适应从共享特征中过滤出特定特征。在对共享特征进行自适应时，使用三个全连接层的模态鉴别器来对每个共享特征的模态进行分类：

$$L_{ma} = -\frac{1}{n}\sum_{i=1}^{n} \log(p(m|H_i^m, \Theta_D)) \tag{4.20}$$

其中 Θ_D 表示模态鉴别器的参数，$p(m|H_i^m, \Theta_D)$ 表示特征 H_i^m 属于模态 m 的预测概率。为了使特定特征与共享特征不相关，使用项目对抗策略，在训练阶段，将特定的特征投射到同一样本的共享特征上。用投影误差作为损失函数：

$$L_{pa} = \frac{1}{n}\sum_{i=1}^{n} \left\| \Theta_P^m \bullet P_i^m - H_i^m \right\| \tag{4.21}$$

其中 Θ_P^m 表示模态 m 的投影矩阵。在识别阶段，Θ_P^m 将特定的特征投射到相应的共享特征中。而在生成阶段，骨干网络会产生与共享特征不相关的特定特征，以欺骗投影矩阵。这种对抗训练可以使两种特征的特征空间线性独立。

为了增强这两个特征的互补性，使用每个模态特征后的解码器网络来重建输入。首先连接共享特征和特定特性，然后将它们提供给解码器 De：

$$\hat{X}^m = De^m([P^m; H^m]) \tag{4.22}$$

使用 L2 损失来评价重建图像的质量：

$$L_{re} = \frac{1}{n}\sum_{i=1}^{n} L_2(X_i^m, \hat{X}_i^m) \tag{4.23}$$

以上损失函数主要对特征提取网络进行训练。使得特定特征和共享特

征能够更有区分度，两类特征在具有自适应性的同时也能互补。

在对特征融合网络进行训练时，需要对融合后的特征进行分类，此处使用分类损失和三元组损失来对网络进行训练。如公式 4.24 和 4.25 所示：

$$L_c(T^m) = -\frac{1}{n}\sum_{i=1}^{n}\log(p(y_i^m \mid \mathrm{T}_i^m)) \tag{4.24}$$

$$\begin{aligned} L_t(T) &= L_{cmT}(T) + L_{smT}(T) \\ &= \sum_{i,j,k}\max[\rho_1 + \left\|T_i^R, T_j^I\right\| - \left\|T_i^R, T_k^I\right\|, 0] \\ &+ \sum_{i,j,k}\max[\rho_1 + \left\|T_i^I, T_j^R\right\| - \left\|T_i^I, T_k^R\right\|, 0] \\ &+ \sum_{i,j,k}\max[\rho_2 + \left\|T_i^R, T_j^R\right\| - \left\|T_i^R, T_k^R\right\|, 0] \\ &+ \sum_{i,j,k}\max[\rho_2 + \left\|T_i^I, T_j^I\right\| - \left\|T_i^I, T_k^I\right\|, 0] \end{aligned} \tag{4.25}$$

4.5 实验及分析

本小节主要介绍了使用非局部注意力网络对双端可共享的行人重识别网络模型进行优化后的相关实验。主要包括实验相关设置、实验环境和实验结果与分析三部分内容。

4.5.1 实验相关设置

本文选取上文中介绍的两种双端网络模型：MACE 和 cm-SSFT，并使用非局部注意力网络对两种模型进行优化。分别对两种模型进行实验，并对两种模型在优化前后的实验结果进行比较。

实验数据集使用 SYSU-MM01 数据集和 RegDB 数据集，在进行网络训练时使用第 3 章提出的数据增强方法分别对两个数据集进行增强。对于 SYSU-MM01 数据集，分别使用了室内搜索模式和全搜索模式进行了实验。对于 RegDB 数据集，使用 3.2.2 介绍的方法进行实验。在每次训练中随机选取 P=8 个身份标签，然后在数据集中随机选取对应身份的 K=4 个可见光图像及 K=4 个红外图像，即每个批次训练包含 32 张可见光图像和 32 张红外图像，总的训练批次大小为 64。训练迭代次数为 60，学习率在前 10 次迭代中由 0.01 递增到 0.1，在第 10 到第 30 次迭代中保持为 0.1，30 次以

后为 0.01。

4.5.2 实验环境

本章实验的实验软硬件环境如表 4.1 所示。

表 4.1 软硬件环境

Table 4.1 Software and hardware environment

硬件	CPU	Inter Core i7-8700 CPU @ 3.2GHz
	GPU	NVIDIA RTX 2080Ti 11GB
	内存	16G
	硬盘	1TB
软件	操作系统	64 位 Ubuntu 16.04 系统
	开发语言	Python 3.6
	开发框架	Pytorch 1.0.1
	开发工具	PyCharm 2019.03

4.5.3 实验结果与分析

本小节给出了在基于第 3 章的基础上，使用非局部注意力网络对 MACE 网络模型和 cm-SSFT 网络模型进行改进前和改进后的实验结果。使用 rank-k 和 mAP 作为评价标准。表 4.2 和表 4.3 分别为两种模型在 SYSU-MM01 数据集的两种模式下的实验结果。

表 4.2 SYSU-MM01 数据集 All-Search 模式下的实验结果

Table 4.2 Experimental results of SYSU-MM01 dataset in all-search mode

Model	All-Search							
	Single-shot				Multi-shot			
	r1	r10	r20	mAP	r1	r10	r20	mAP
MACE	43.6	76.5	86.3	54.1	47.1	83.7	90.3	47.2
MACE+Non-local	44.9	77.9	87.2	55.3	48.8	84.2	93.1	48.5
cm-SSFT	58.4	86.7	89.9	61.4	59.6	88.4	92.9	59.6
cm-SSFT+Non-local	59.3	87.5	91.7	62.1	61.1	90.1	93.5	61.6

表 4.3 SYSU-MM01 数据集 Indoor-Search 模式下的实验结果

Table 4.3 Experimental results of SYSU-MM01 dataset in Indoor-search mode

Model	Indoor-Search							
	Single-shot				Multi-shot			
	r1	r10	r20	mAP	r1	r10	r20	mAP
MACE	46.5	78.7	88.4	63.1	53.9	88.7	94.6	61.4
MACE+Non-local	47.7	80.3	90.5	64.1	55.4	90.2	95.6	62.9

cm-SSFT	66.3	90.6	92.8	69.3	69.1	92.5	95.9	69.1
cm-SSFT+Non-local	68.3	92.1	94.0	71.3	71.7	94.1	97.1	71.9

表 4.4 记录了上述两种模型在 RegDB 数据集的实验结果，其中包含了 Visible to Thermal 和 Thermal to Visible 两种实验方法。

表 4.4 RegDB 数据集实验结果

Table 4.4 Experimental results of RegDB dataset

Model	Visible to Thermal		Thermal to Visible	
	r1	mAP	r1	mAP
MACE	68.6	67.1	67.3	65.8
MACE+Non-local	70.7	68.3	68.8	66.7
cm-SSFT	69.8	69.9	67.3	68.9
cm-SSFT+Non-local	71.2	71.6	68.6	70.3

由上述实验可知，使用非局部注意力网络对两种模型进行优化，rank-k 和 mAP 均有所提升。同时因为对数据进行了数据增强，所以整体的识别准确率仍然要低于不使用数据增强时的识别准确率。尽管数据增强对于双端网络模型增加了识别的难度，但非局部注意力网络嵌入特征提取网络后，其感受野扩大到整个特征图的大小，对于空间上存在距离的相关联特征来说，能够同时被非局部注意力网络感受到，类似于被同一个卷积核捕获，非局部注意力网络会把相关联的特征进行捕获，使得两者有了更大的相互融合的概率，进一步增强了双端网络模型的表征学习能力。

为了更直观的体现非局部注意力网络对提升行人重识别模型识别准确率的影响，本章给出了可视化的结果。使用 MACE 算法模型为 Baseline，使用 SYSU-MM01 数据集作为实验数据集，对其使用非局部注意力网络前后的实验结果进行可视化。如图 4.6 所示，图中展示了排序队列中前 20 张图像，该 20 张图像根据与基准样本的相似度由大到小排列，其中绿色框线的图像为预测正确的图像，即正样本。由图可知，使用非局部注意力网络对模型进行优化后，一些比较难的正样本其最终的排序结果相对优化之前其排序位置更靠前，意味着网络对其预测的更准确，同时使用非局部注意力网络优化后的网络模型整体的预测准确率也更高。该结果直观的证明了非局部注意力网络对于提高网络重识别的准确率有着积极的作用。

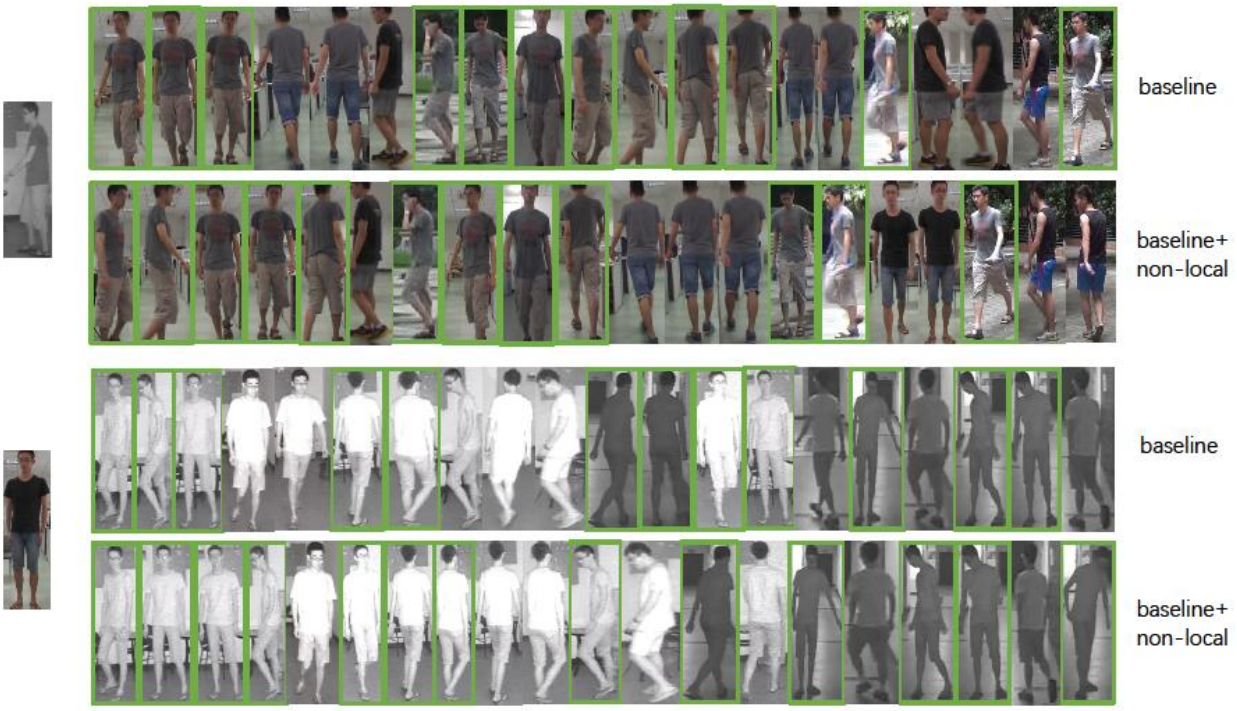


图 4.6 多模态行人重识别可视化结果

Fig 4.6 Visualization results of multimodal pedestrian re-recognition

非局部注意力网络捕获了共享特征中的长距离依赖，从而提高了共享特征提取网络的特征提取效率，也使得共享网络提取到了更丰富的共享特征。以上的实验结果证实了非局部注意力网络对于共享特征提取能力的正面促进作用，由此也证明了非局部注意力网络对于多模态下双端可共享网络的行人重识别网络模型的优化是可行的。

4.6 本章小结

本章主要针对多模态行人重识别任务中，在对共享特征进行提取时无法捕获长距离依赖的问题进行了分析，并对该问题提出了优化方案。本章首先指出了在当前的网络模型中存在的问题，即对于融合后的特征，其内部存在空间距离上的长距离依赖，导致一般的卷积神经网络无法有效的提取这种长距离依赖。随后本章提出了使用非局部注意力网络来对特征提取网络进行改进，由于非局部注意力网络可以将感受野扩大到整个特征图，所以可以很容易的捕获到长距离依赖。其次分别介绍了 MACE 和 cm-SSFT 两种双端网络模型，同时使用非局部注意力网络对这两种双端网络模型进

行改进，将非局部注意力网络嵌入其共享特征提取网络中，用以捕获长距离依赖。最后对改进后的网络模型进行了实验，实验结果表明，经过优化后的网络其 rank-k 和 mAP 均有所提升，从而证明了非局部注意力网络对于提高共享特征提取能力的有效性。

第 5 章 基于特征均值聚类损失的多模态行人重识别模型

在本文第 4 章中，通过引入非局部注意力网络来对多模态下行人重识别任务的表征学习进行了改进，提高了网络的特征提取能力。在多模态行人重识别的任务中，除了对网络进行表征学习，还需要使用度量学习来对网络进行训练。一般使用三元组损失进行度量学习，同时为了提高训练效率，大部分网络模型都会进行难样本挖掘。但使用难样本挖掘的三元组损失只会对基准样本与难正样本和难负样本之间的特征距离进行优化，使得样本空间中仍然存在较大的类内距离以及较小的类间距离，因此三元组损失对于多模态行人重识别任务并不理想。本章节中使用基于特征均值的聚类损失代替三元组损失进行度量学习，清除了较大的类内距离和较小的类间距离，使之能够将相同标签的样本进行聚类，并使用优化后的网络模型进行了实验，最后对实验结果进行了分析。

5.1 问题提出

在行人重识别任务中，大部分模型都通过表征学习和度量学习结合的方式来训练网络，如：在表征学习中，将重识别任务当作分类任务，使用交叉熵损失对分类器网络进行优化；在度量学习中，学习两张图片之间的相似度，并使用三元组损失、对比损失等来进行网络训练。在行人重识别任务中，度量学习的主要表现为：同一行人的不同图片的相似度大于不同行人不同图片之间的相似度。多模态行人重识别任务中，进行度量学习时主要通过三元组损失来实现，并且为了提高训练效率，一般选择具有难样本挖掘的三元组损失。使用 a、p、n 分别代表基准样本、正样本和负样本。在难样本挖掘的三元组损失中，需要在所有的正样本 p 中找到距离基准样本 n 最远的难正样本，同时还需要在所有的负样本 n 中找到距离基准样本 n 最近的难负样本，使得基准样本与难正样本之间的距离小于基准样本与难负样本之间的距离，即：$\max d_{a,p} < \min d_{a,n}$，以此来达到聚类的效果。

使用难样本挖掘可以缩小相同标签下样本间的最大差异，同时扩大不

同标签下样本间的最小差异。但三元组损失在多模态行人重识别任务中表现得效果并不理想，主要原因在于三元组损失中使用的难样本挖掘分别只选择了一个正样本和一个负样本，其它样本与基准样本之间的距离并没有发生变化，并且也没有对度量学习产生直接或间接的贡献，所以在使用难样本挖掘后，相同标签的样本中仍然存在较大的类内差距。如图 5.1 所示，通过难样本挖掘的三元组损失，仅仅优化了距离最远的正样本和距离最近的负样本，而对于距离相对较大的正样本和距离相对较小的负样本并没有进行聚类和优化，这就导致样本空间内仍然存在较大的类内差异和较小的类间差异，从而使得相同类别的样本不能很好的聚类到一起。

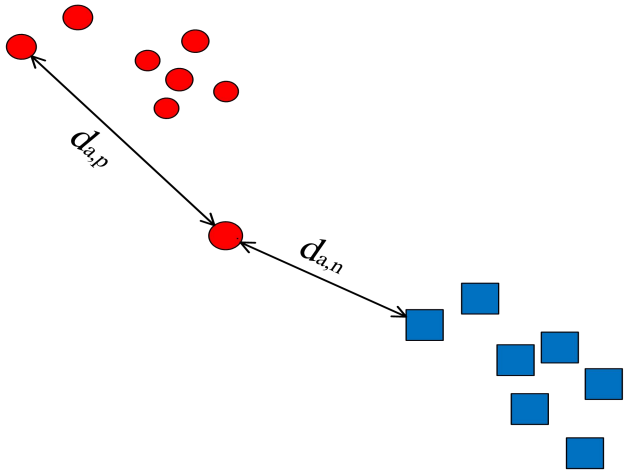


图 5.1 三元组损失示意图
Fig 5.1 Triplet loss diagram

5.2 多模态下基于特征均值的聚类损失函数

本章使用基于特征均值的聚类损失函数来代替难样本挖掘的三元组损失函数。使得损失函数不仅优化难样本之间的距离，还间接地优化所有类内图像之间基于均值的距离，从而使训练批次中的所有样本都对损失函数有贡献。

在多模态行人重识别的任务中，使用基于特征均值的聚类损失可以有效的减小基准样本与各模态下正样本的特征距离，同时扩大与各模态下负样本的特征距离。由于每个身份的图像同时包含可将该图像和红外图像，所以此处设每个身份有 K 个可见光图像和 K 个红外图像，则整个训练过程的 Batch size 大小为 $2K\times P$。

5.2.1 基于特征均值聚类损失的 MACE 算法模型

在 4.3 介绍的 MACE 算法模型中，多模态图像在进入共享特征提取网络后，会得到两种模态图像的共享特征，记作 H。由于该共享特征来自于两种模态下的两张图像，且在进入共享网络之前的输入是由两种模态的特征拼接而成，因此可以对该共享特征进行拆分，拆分后将所得的特征分别记为 H^V 和 H^I，分别表示共享网络从两种模态的图像中提取的共享特征。则某一身份 i 在两种模态下的特征均值可分别表示为：

$$H_i^{Vm} = \frac{\sum_K H^V}{K} \tag{5.1}$$

$$H_i^{Im} = \frac{\sum_K H^I}{K} \tag{5.2}$$

身份 i 对应的类内距离由该身份的每个样本特征到该身份的特征均值的距离表示，同时为了提高训练效率进行难样本挖掘，所以选取所有距离之间的最大值。则特定模态内的类内距离 $d_i^{intra(V-V)}$ 和 $d_i^{intra(I-I)}$ 计算如公式 5.3 和 5.4 所示：

$$d_i^{intra(V-V)} = \max_K \left\| H^V - H_i^{Vm} \right\|_2^2 \tag{5.3}$$

$$d_i^{intra(I-I)} = \max_K \left\| H^I - H_i^{Im} \right\|_2^2 \tag{5.4}$$

同时考虑到跨模态下的场景，不同模态下的同一身份的图片特征同样需要进行聚类。因为基准样本可以是可见光图像，也可以是红外图像，所以此处需包含两种类型的类内距离，即 $d_i^{intra(V-I)}$ 和 $d_i^{intra(I-V)}$，计算公式如式 5.5 和式 5.6 所示：

$$d_i^{intra(V-I)} = \max_K \left\| H^V - H_i^{\mathrm{Im}} \right\|_2^2 \tag{5.5}$$

$$d_i^{intra(I-V)} = \max_K \left\| H^I - H_i^{V\mathrm{m}} \right\|_2^2 \tag{5.6}$$

同样地，一个身份的类间距离由该身份的特征均值到同一批次中所有其他身份的特征均值的距离来表示，为了挖掘难样本，此处选择距离最小值。则特定模态内的类间距离 $d_i^{inter(V-V)}$ 和 $d_i^{inter(I-I)}$ 的计算公式如式 5.7 和式 5.8 所示：

$$d_i^{inter(V-V)} = \min_{\forall id \in P, id \neq i} \left\| H_i^{Vm} - H_{id}^{Vm} \right\|_2^2 \tag{5.7}$$

$$d_i^{inter(I-I)} = \min_{\forall id \in P, id \neq i} \left\| H_i^{Im} - H_{id}^{Im} \right\|_2^2 \tag{5.8}$$

与类内距离类似，不同模态下的不同身份的图片特征需要扩大其特征距离，此时，跨模态下的类间距离由该身份的特征均值到同一批次内不同模态下其它身份的特征均值的距离来表示，同样选取最小值来进行难样本挖掘，则跨模态下的类间距离 $d_i^{inter(V-I)}$ 和 $d_i^{inter(I-V)}$ 计算公式如式 5.9 和式 5.10 所示：

$$d_i^{inter(V-I)} = \min_{\forall id \in P, id \neq i} \left\| H_i^{Vm} - H_{id}^{\mathrm{Im}} \right\|_2^2 \tag{5.9}$$

$$d_i^{inter(I-V)} = \min_{\forall id \in P, id \neq i} \left\| H_i^{Im} - H_{id}^{V\mathrm{m}} \right\|_2^2 \tag{5.10}$$

由以上计算公式分别可求得 4 个类间距离和 4 个类内距离，因此可根据以上的两种距离得到 4 个聚类损失函数，即特定模态下的 L_C^{V-V}、L_C^{I-I} 和跨模态下的 L_C^{V-I}、L_C^{I-V}，计算公式如下所示：

$$L_C^{V-V} = \sum_{I}^{P} \max((d_i^{intra(V-V)} - d_i^{inter(V-V)} + \alpha), 0) \tag{5.11}$$

$$L_C^{I-I} = \sum_{I}^{P} \max((d_i^{intra(I-I)} - d_i^{inter(I-I)} + \alpha), 0) \tag{5.12}$$

$$L_C^{V-I} = \sum_{I}^{P} \max((d_i^{intra(V-I)} - d_i^{inter(V-I)} + \alpha), 0) \tag{5.13}$$

$$L_C^{I-V} = \sum_{I}^{P} \max((d_i^{intra(I-V)} - d_i^{inter(I-V)} + \alpha), 0) \tag{5.14}$$

其中α为目标样本分别与类内和类间距离的间隔参数，设为 0.2。由于聚类损失函数是基于类间距离和类内距离之差计算的，必须保证四个函数中的差值有一个全局性的最小值，因此四个函数采用相同的α 值。

考虑到实际应用中跨模态或多模态任务是随机且均匀的，所以本文认为以上四个聚类损失函数对总的聚类损失函数有相同的贡献。最终得到总的聚类损失为：

$$L_C = L_C^{V-V} + L_C^{I-I} + L_C^{V-I} + L_C^{I-V} \tag{5.15}$$

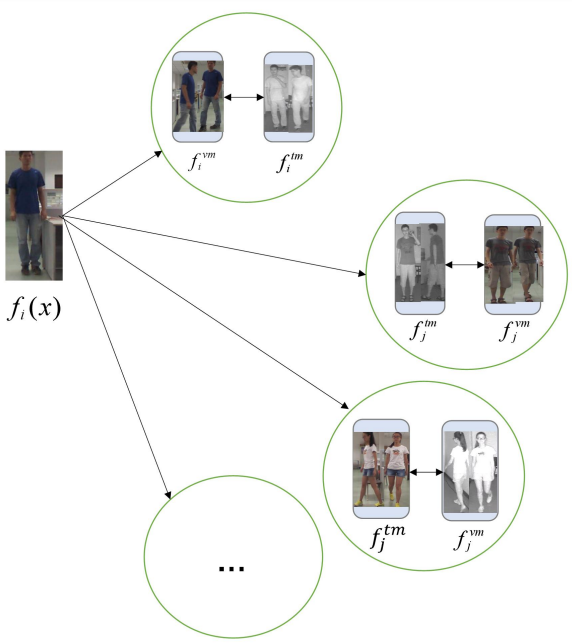


图 5.2 多模态聚类示意图
Fig 5.2 Schematic diagram of multimodal clustering

多模态下的聚类损失示意图如图 5.2 所示。使用基于特征均值的聚类损失函数对网络进行训练，不仅可以对难样本进行挖掘，并且由于类内距离和类间距离是通过各模态的特征均值计算所得的，所以使得每个样本都间接地对网络训练作出贡献，解决了三元组损失中依然存在较大类内距离的问题。

5.2.2 基于特征均值聚类损失的 cm-SSFT 算法模型

在 cm-SSFT 算法模型中，需要同时将训练特征提取网络的三元组损失函数和训练特征融合网络的三元组损失函数替换为聚类损失函数，如图 4.5 所示。在特征提取阶段，cm-SSFT 算法不仅关注各模态图像的共享特征，也关注每个模态的特定特征，因此在对共享特征进行分解后会得到四个特征，除了共享特征 H^V 和 H^I，还包括可将光模态的特定特征 P^V 和红外光模态的特定特征 P^I。对于 H^V 和 H^I，其基于均值的聚类损失与 5.2.1 中的计算方式一致，本小节不再列出。对于特定模态下的特征，其计算方式可视为单模态下的聚类损失来进行计算。

在两种特定模态中，对于身份 i，各模态下的特征均值分别表示为：

$$P_i^{Vm} = \frac{\sum^K P^V}{K} \tag{5.16}$$

$$P_i^{Im} = \frac{\sum^K P^I}{K} \tag{5.17}$$

通过计算每个模态下样本特征到特征均值之间的距离来计算各模态下的类内距离 d_i^{intra} 和类间距离 d_i^{inter}，同时进行难样本挖掘，计算公式如下所示：

$$d_i^{intra(V)} = \max_K \left\| P^V - P_i^{Vm} \right\|_2^2 \tag{5.18}$$

$$d_i^{inter(V)} = \max_{\forall id \in P, id \neq i} \left\| P_{id}^{Vm} - P_i^{Vm} \right\|_2^2 \tag{5.19}$$

同理，对于红外光模态下的类内距离和类间距离计算如下：

$$d_i^{intra(I)} = \max_K \left\| P^I - P_i^{Im} \right\|_2^2 \tag{5.20}$$

$$d_i^{inter(I)} = \max_{\forall id \in P, id \neq i} \left\| P_{id}^{Im} - P_i^{Im} \right\|_2^2 \tag{5.21}$$

则单模态下的聚类损失函数分别为：

$$L_c^V = \sum_i^P \max((d_i^{intra(V)} - d_i^{inter(V)} + \alpha), 0) \tag{5.22}$$

$$L_c^I = \sum_i^P \max((d_i^{intra(I)} - d_i^{inter(I)} + \alpha), 0) \tag{5.23}$$

使用 L_c^H 表示使用共享特征计算得到的聚类损失，则对于特征提取网络，其基于特征均值的聚类损失 L_{c1} 如下式所示：

$$L_{c1} = L_c^V + L_c^i \tag{5.24}$$

对于特征融合网络的聚类损失函数，在特征学习网络中获得共享特征 T 后，需要对 T 进行分解，得到两个共享特征，分别是来自可见光模态的共享特征 H^V 和来自红外光模态的共享特征 H^I，使用公式 5.3- 5.6 计算来获得特定模态内的类内距离 $d_i^{intra(V-V)}$、$d_i^{intra(I-I)}$ 和跨模态的类内距离 $d_i^{intra(V-I)}$、$d_i^{intra(I-V)}$，使用公式 5.7-5.10 计算获得特定模态内的类间距离 $d_i^{inter(V-V)}$、$d_i^{inter(I-I)}$ 和跨模态下的类间距离 $d_i^{inter(V-I)}$ 和 $d_i^{inter(I-V)}$，然后使用公式 5.11-5.14 计算获得四种类型的聚类损失。最后根据所求得的聚类损失来对特征学习网络进行训练。

5.3 实验及分析

本小节在基于第 4 章优化的基础上对 MACE 算法模型和 cm-SSFT 算法模型进行了改进，对使用基于特征均值的聚类损失优化后的网络模型进行了实验。主要包括实验相关设置、实验环境和实验结果与分析三部分内容。

5.3.1 实验相关设置

本章实验主要在第 4 章的基础上，使用 5.2 中提出的基于特征均值的聚类损失函数，分别将 MACE 算法模型和 cm-SSFT 算法模型中的三元组损失函数替换为基于特征均值的聚类损失函数，并对改进后的两种模型进行了实验。

实验数据集使用 SYSU-MM01 数据集和 RegDB 数据集。对于 SYSU-MM01 数据集，分别使用了室内搜索模式和全搜索模式进行了实验。对于 RegDB 数据集，使用 3.2.2 介绍的方法进行实验。在每次训练中随机选取 P=8 个身份标签，然后在数据集中随机选取对应身份的 K=4 个可见光图像及 K=4 个红外图像，即每个批次训练包含 32 张可见光图像和 32 张红外图像，总的训练批次大小为 64。训练迭代次数为 60，学习率在前 10 次迭代中由 0.01 递增到 0.1，在第 10 到第 30 次迭代中保持为 0.1，30 次以后为 0.01。

同时还对两类损失函数进行了聚类结果可视化的对比试验，对比实验中使用 Resnet-50 作为特征提取网络，并使用 SYSU-MM01 数据集中 cam1 和 cam3 下的前十个行人身份的可见光图像和红外图像作为输入数据。

5.5.2 实验环境

本章实验的实验软硬件环境如表 5.1 所示。

表 5.1 软硬件环境

Table 5.1 Software and hardware environment

硬件	CPU	Inter Core i7-8700 CPU @ 3.2GHz
	GPU	NVIDIA RTX 2080Ti 11GB
	内存	16G
	硬盘	1TB
软件	操作系统	64 位 Ubuntu 16.04 系统

开发语言	Python 3.6
开发框架	Pytorch 1.0.1
开发工具	PyCharm 2019.03

5.5.3 实验结果与分析

本小节给出了使用基于特征均值的聚类损失函数对 MACE 算法模型和 cm-SSFT 算法模型优化后的实验结果,其中包括单独使用聚类损失函数优化和同时使用非局部注意力网络与聚类损失函数优化两种情况。实验使用 rank-k 和 mAP 作为评价标准。表 5.2 和表 5.3 分别为两种模型在 SYSU-MM01 数据集的两种模式下的实验结果，同时为了方便与优化前的网络模型进行比较，表中列出了在没有使用本文提出的所以优化方法的原网络模型的实验结果。

表 5.2 SYSU-MM01 数据集 All-Search 模式下的实验结果

Table 5.2 Experimental results of SYSU-MM01 dataset in all-search mode

Model	All-Search							
	Single-shot				Multi-shot			
	r1	r10	r20	mAP	r1	r10	r20	mAP
MACE	45.3	79.8	91.1	56.9	51.6	87.3	94.4	50.1
MACE+Cluster-loss	45.7	80.4	91.8	57.4	51.8	87.7	95.0	50.6
MACE+Non-local+Cluster-loss	46.8	81.2	92.6	59.5	52.4	89.1	96.9	52.7
cm-SSFT	61.6	89.2	93.9	63.2	63.4	91.2	95.7	62.0
cm-SSFT+Cluster-loss	62.1	89.5	94.7	64.1	63.6	91.7	95.8	63.1
cm-SSFT+Non-local+Cluster-loss	63.9	91.4	96.6	67.0	65.7	92.3	97.2	64.4

表 5.3 SYSU-MM01 数据集 Indoor-Search 模式下的实验结果

Table 5.3 Experimental results of SYSU-MM01 dataset in Indoor-search mode

Model	Indoor-Search							
	Single-shot				Multi-shot			
	r1	r10	r20	mAP	r1	r10	r20	mAP
MACE	48.3	81.8	91.6	66.2	57.4	93.0	97.5	64.8
MACE+Cluster-loss	48.6	79.9	89.8	64.2	55.1	90.4	95.5	63.2
MACE+Non-local+Cluster-loss	49.4	82.4	91.9	66.5	57.6	93.7	97.7	65.3
cm-SSFT	70.5	94.9	97.7	72.6	73.0	96.3	99.1	72.4
cm-SSFT+Cluster-loss	70.7	92.8	94.1	71.8	71.2	94.8	96.7	71.6
cm-SSFT+Non-local+Cluster-loss	71.2	95.2	98.6	73.2	73.9	96.9	99.4	73.9

表 5.4 记录了上述两种模型在 RegDB 数据集的实验结果，其中包含了 Visible to Thermal 和 Thermal to Visible 两种实验方法。

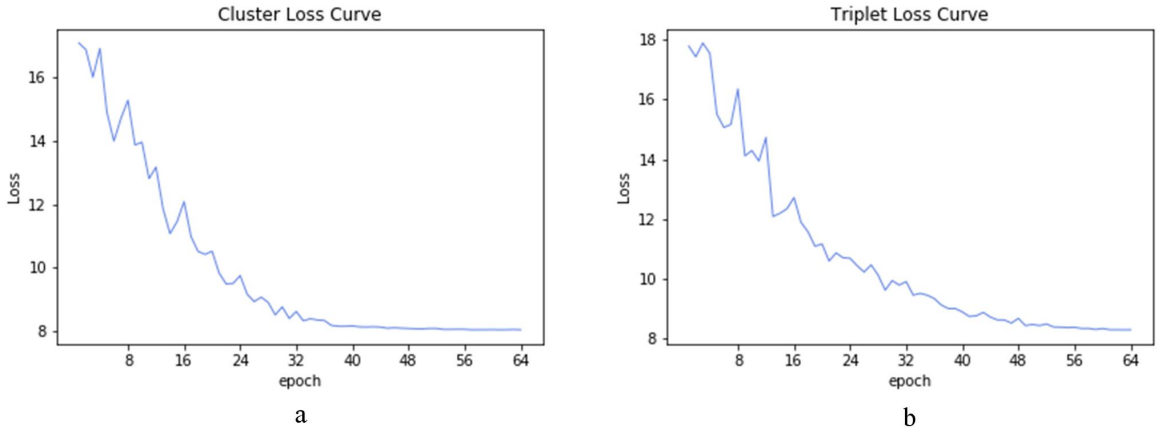
表 5.4 RegDB 数据集实验结果

Table 5.4 Experimental results of RegDB dataset

Model	Visible to Thermal		Thermal to Visible	
	r1	mAP	r1	mAP
MACE	72.4	69.1	72.1	68.6
MACE+Cluster-loss	69.8	68.4	69.5	67.3
MACE+Non-local+Cluster-loss	72.9	71.3	72.4	69.7
cm-SSFT	72.3	72.9	71.0	71.7
cm-SSFT+Cluster-loss	72.8	73.3	71.2	71.9
cm-SSFT+Non-local+Cluster-loss	73.5	73.7	71.8	72.4

由上述实验可知，使用基于特征均值的聚类损失函数对上述两种双端网络模型进行优化，其 rank-k 和 mAP 均有所提升，并且分别在没有非局部注意力网络优化和有局部注意力网络优化的基础上进行了实验，实验结果均表明，基于特征均值的聚类损失函数对于多模态下的行人重识别网络模型的性能提高有积极的作用。

为便于比较三元组损失和基于特征均值的聚类损失两者的收敛速度，分别使用两种损失函数对 MACE 算法进行优化，并分别在 SYSU-MM01 数据集的 All-Search（Single-shot）模式和 Indoor-Search（Single-shot）模式下进行实验，同时实验过程中分别记录了网络模型各自对应的损失函数在训练中每个 batch 的数值，图 5.3 和图 5.4 展示了训练过程中两张损失函数的变化情况。



a　　b

图 5.3 All-Search 模式特征均值聚类损失（a）和三元组损失（b）

Fig 5.3 Mean feature cluster loss (a) and Triplet loss (b) in All - Search mode

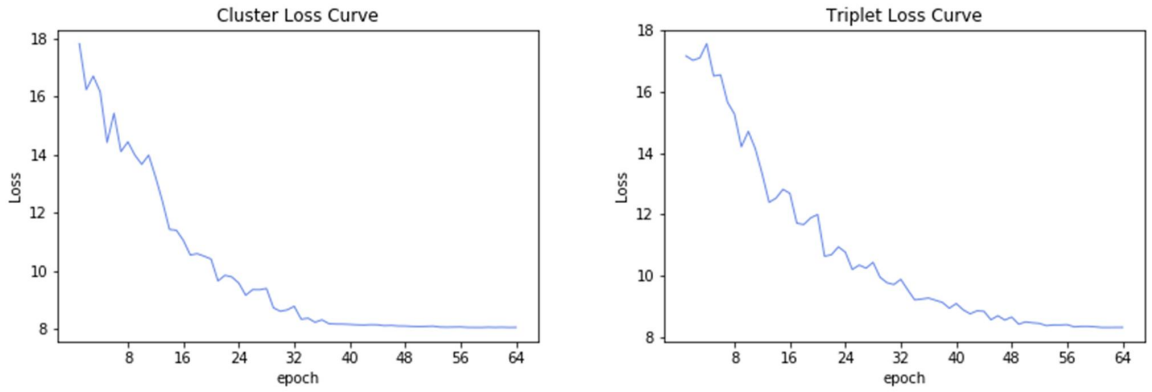


图 5.4 Indoor-Search 模式特征均值聚类损失（a）和三元组损失（b）
Fig 5.4 Mean feature cluster loss (a) and Triplet loss (b) in Indoor - Search mode

图 5.3 和图 5.4 显示基于特征均值的聚类损失函数在训练过程中其收敛速度比三元组损失函数的收敛速度更快，聚类损失函数在第 40 批次时开始收敛，而三元组损失函数在第 56 批次左右时开始收敛。MACE 算法模型和 cm-SSFT 算法模型作为双端共享网络模型，其本身对于共享特征及特定特征的区分度比较敏感，如果网络中无法有效的对两者模态的特征进行区分，则对于两者的共享特征网络。其对输入图像的共享特征的提取能力便会受到影响。基于特征均值的聚类损失函数对不同类别的特征进行聚类，由于是基于特征均值的聚类函数，所以间接的使每一个样本都参与了聚类，使得聚类结果更有区分度。尤其是对于 cm-SSFT 算法模型，该模型除了要提取共享特征，还要提取各个模态的特定特征用以辅助分类器进行识别。样本空间中不同类的样本区分度增大，同类样本之间距离更近，对于特定模态的特征提取器，更容易提取到区分度更大的特征。

为了更直观地观察聚类损失函数与三元组损失函数之间的区别，本文对两者之间的聚类效果进行了可视化。使用 ResNet-50 作为骨干网络，并采用图 4.1 所示的双端网络结构，同时以 SYSU-MM01 数据集中 cam1 和 cam3 下的前十个行人身份的可见光图像和红外图像作为输入数据，聚类结果经过可视化后如图 5.5 所示。由图可知，基于特征均值的聚类损失函数在对数据进行聚类后，其类内距离相比于三元组损失函数聚类后的类内距离更均匀，并且不同类之间的类间距离也更大。三元组损失对样本聚类后在同一类的样本空间中还存在较大的类内距离，同一类的样本空间中，样本分布也相比基于特征均值的聚类损失的样本分布更分散。同时三元组损

失函数在聚类后不同类之间的类间距离相对较小。对于上述现象，分析原因为：由于聚类损失

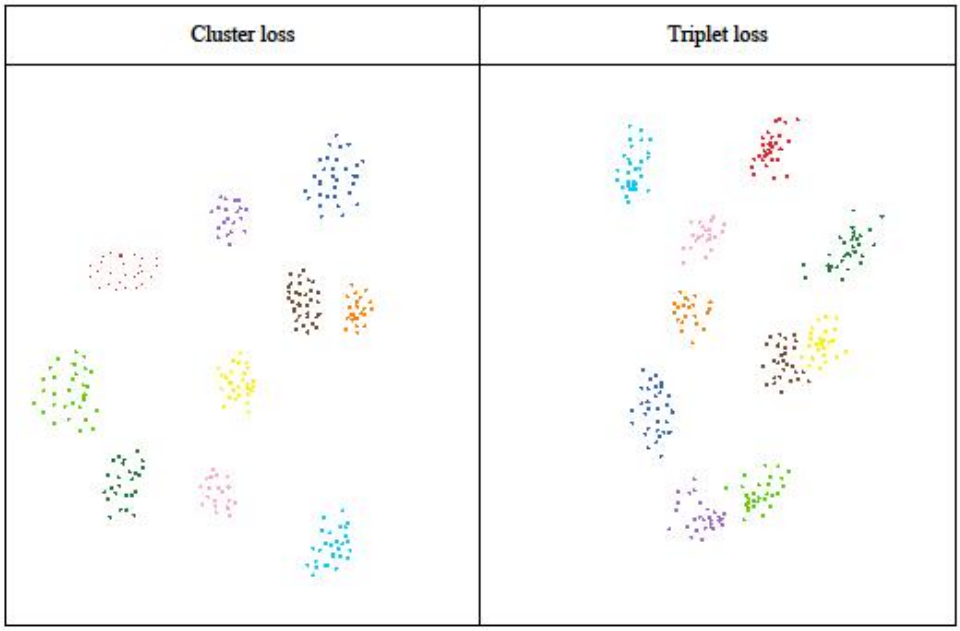


图 5.5 聚类损失与三元组损失对比可视化
Fig 5.5 Comparison visualization of cluster loss and triplet loss

函数是基于特征均值的，使得在基于特征均值的聚类损失函数在进行优化的过程中，每一个样本都间接的参与了优化，也使得类内距离更均匀，不会存在较大的类内距离，同时基于特征均值的聚类过程会使不同类之间更有区分度，从而使类间距离增大。

基于特征均值的聚类损失函数使用每个特征相对于特征均值的相对距离进行优化，由于有特征均值的参与，使得在优化过程中每个样本也在间接的参与优化，这就使得聚类效果比三元组损失更加明显。三元组损失在进行难样本挖掘后，样本空间中依然存在较大的类内距离和较小的类间距离，这就导致网络的度量学习还有改进的空间，基于特征均值的聚类损失进一步优化了特征距离。本章实验也验证了基于特征均值的聚类损失函数的有效性。

最后，相较于优化前的两种原网络模型，在使用数据增强的基础上，使用非局部注意力网络和基于特征均值的聚类损失函数对两个原有的网络模型进行优化后，两者的 rank-k 和 mAP 也都有所提升。非局部注意力网络对行人重识别网络的表征学习进行了增强，基于特征均值的聚类损失函数对行人重识别网络的度量学习进行了优化，使得网络不仅提高了特征的

提取能力，同时也提高了特征之间相似度的学习能力，最终使得网络处理多模态行人重识别任务的准确率提升。

5.4 本章小结

本章主要针对以往的多模态行人重识别网络中使用三元组损失进行度量学习时存在的问题进行了优化。三元组损失在进行难样本挖掘后，样本空间中依然存在较大的类内距离和较小的类间距离，导致聚类效果不明显，本小节使用了基于特征均值的聚类损失函数来替代三元组损失函数，同时在基于第 4 章的基础上对 MACE 和 cm-SSFT 两种网络模型进行了优化，最后对改进后的两种网络进行了实验。实验结果表明，经过优化后的网络其 rank-k 和 mAP 均有所提升，从而证明了基于特征均值的聚类损失函数在改进度量学习、提升网络的重识别准确率等方面的有效性。

第 6 章 总结与展望

近年来，由于国家的大力支持，监控系统越来越普及，政府推出了一系列措施，如智慧城市、平安城市、雪亮工程等，使得公共空间的监控摄像头密度较之前有了大幅增长。监控系统的普及意味着数据的增加，对于目标的搜索增加了难度，同时为了捕获光照条件不足时的影像，红外摄像头也越来越多的被使用，这也使得有大量的监控图像为多模态数据，这也为目标搜寻增加了难度，多模态行人重识别技术正是基于以上的现实需求应运而生的。随着深度学习技术的发展，越来越多的深度神经网络被应用于计算机视觉领域，在多模态行人重识别领域也涌现出了很多优秀的深度神经网络模型，双端可共享网络模型是其中比较常用的一种模型。以往的双端网络在表征学习中没有关注数据中存在的行人姿态不对齐问题，因此导致训练出的模型可能存在一定的过拟合。同时在进行特征融合时没有关注长距离依赖的捕获，因此网络的特征提取能力受到一定的限制。此外，行人重识别网络中应用较多的三元组损失并没有很好的对样本进行聚类，导致样本空间中仍然存在较大的类内距离和较小的类间距离，使得网络训练的效果不佳，直接影响重识别网络的识别准确率。本文对以上的这些问题分别做出了改进和优化，并取得了一定的效果。本章对本文的工作进行了总结，并结合当前在该领域中研究的不足对未来的研究进行了展望。

6.1 本文主要工作

本文主要针对多模态行人重识别任务中的双端可共享网络模型存在的一些问题进行了分析和改进，主要完成了一下工作：

（1）提出了一种基于行人姿态不对齐的行人重识别数据增强方法

由于行人重识别任务中的数据来自于监控图像，并且由图像识别算法从中截取的，因此会存在大量的行人姿态不对齐的现象。而训练网络的数据集中其行人图像大多都是姿态对齐的，所以训练所得的网络对于真实数据存在一定的过拟合现象，并没有很好的鲁棒性。本文提出了一种模拟行人姿态不对齐的数据增强方法，选择每个行人身份所对应图片数量的四分之一作为数据增强的数据，将调整后的图像与原数据一起作为训练数据。

经过实验，该数据增强方法减少了网络模型的过拟合现象，使网络具有一定的鲁棒性。

（2）提出了一种基于非局部注意力网络的多模态行人重识别算法

多模态下的双端网络在进行共享特征提取时，会在共享网络之前进行特征融合，然后进行共享特征的提取。但由于融合后的特征图中存在长距离的相关联特征，特征提取网络中的卷积层无法扩大感受野，使得对于长距离依赖的特征提取能力不足，导致无法获得有效的共享特征。本文使用非局部注意力网络嵌入到特征提取网络中，在扩大感受野捕获长距离依赖的同时，也可以保证特征提取网络的输入和输出的大小保持不变。实验表明，非局部注意力网络对于共享特征的提取起到了积极的作用。

（3）提出了一种基于特征均值聚类损失的多模态行人重识别算法

一般的多模态行人重识别网络采用三元组损失来训练网络，尽管对难样本进行了挖掘，但由于没有考虑样本空间中的所有样本，使得样本空间中仍然存在较大的类内距离和较小的类间距离，聚类效果并不明显。本文提出了使用基于特征均值的聚类损失来对网络进行训练，由于使用的距离是样本特征相对于特征均值之间的距离，所以使得所有样本都间接的参与了网络训练。相对于三元组损失，样本空间内不会存在较大的类内距离和较小的类间距离，聚类效果更好。经过实验证明，基于特征均值的聚类损失对于多模态下行人重识别准确率有了很好的提升。

6.2 进一步工作

从实验结果可知，本文提出的优化算法对于行人重识别网络的识别能力有了很大的提高，但相较于传统的图像识别算法，其准确率还有待提升。所以本文的工作还有许多可改进的空间：

（1）在深度学习算法上的改进

本文中使用的深度神经网络主要用来进行特征提取，在深度学习领域有很多优秀的网络框架可用于改进算法，如对抗生成网络（GAN）。可以使用对抗生成网络来生成不同模态的数据，并使用生成的数据来弥补当前模态所对应的不足，这样可以增强网络的学习能力。同时也可以在神经网络中加入相应的模块，来对行人姿态不对齐的问题进行学习，即使用度量学习的方法来解决行人姿态不对齐的问题。

（2）在数据来源上的改进

行人重识别任务中的数据来源于行人识别任务，所以其输入数据的信息量主要由行人识别算法所决定。当前行人重识别任务使用的数据集中，其图像分辨率普遍较低，对于识别算法的识别能力具有很大的限制。因此本文以后的工作可以关注行人识别阶段的算法，可以使用 GAN 技术对图像进行增强，丰富重识别任务中输入数据的信息，从而使得重识别网络可提取到的信息更丰富，进而提高网络的识别准确率。

希望在未来的工作中可以从以上两方面进行改进和完善，使得多模态下的行人重识别任务准确率能够更高。

参考文献

1. Q. Leng, M. Ye, and Q. Tian, “A survey of open-world person reidentification,” IEEE Transactions on Circuits and Systems for Video Technology (TCSVT), 2019.1.
2. L. Zheng, H. Zhang, S. Sun, M. Chandraker, Y. Yang, and Q. Tian, “Person re-identification in the wild,” in CVPR, 2017, pp. 1367–1376
3. Z. Zheng, L. Zheng, and Y. Yang, “Unlabeled samples generated by gan improve the person re-identification baseline in vitro,” in ICCV, 2017, pp. 3754–3762.
4. Z. Zhong, L. Zheng, D. Cao, and S. Li, “Re-ranking person reidentification with k-reciprocal encoding,” in CVPR, 2017, pp.1318–1327.
5. Y. Suh, J. Wang, S. Tang, T. Mei, and K. Mu Lee, “Part-aligned bilinear representations for person re-identification,” in ECCV, 2018, pp. 402–419.
6. L. Zhao, X. Li, Y. Zhuang, and J. Wang, “Deeply-learned partaligned representations for person re-identification,” in ICCV,2017, pp. 3219–3228.
7. D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng, “Person reidentification by multi-channel parts-based CNN with improved triplet loss function,” in CVPR, 2016, pp. 1335–1344.
8. D. Li, X. Chen, Z. Zhang, and K. Huang, “Learning deep context-aware features over body and latent parts for person reidentification,” in CVPR, 2017, pp. 384–393.
9. H. Zhao, M. Tian, S. Sun, and et al, “Spindle net: Person reidentification with human body region guided feature decomposition and fusion,” in CVPR, 2017, pp. 1077–1085.
10. Y. Suh, J. Wang, S. Tang, T. Mei, and K. Mu Lee, “Part-aligned bilinear representations for person re-identification,” in ECCV, 2018, pp. 402–419.

11. Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang, “Beyond part models: Person retrieval with refined part pooling,” in ECCV, 2018, pp. 480–496.
12. C. Su, S. Zhang, J. Xing, W. Gao, and Q. Tian, “Deep attributes driven multi-camera person re-identification,” in ECCV, 2016, pp. 475–491.
13. X. Chang, T. M. Hospedales, and T. Xiang, “Multi-level factorisation net for person re-identification,” in CVPR, 2018, pp. 2109–2118.
14. F. Liu and L. Zhang, “View confusion feature learning for person re-identification,” in ICCV, 2019, pp. 6639–6648.
15. Z. Zhu, X. Jiang, F. Zheng, X. Guo, F. Huang, W. Zheng, and X. Sun, “Aware loss with angular regularization for person reidentification,” in AAAI, 2020.
16. J. Liu, B. Ni, Y. Yan, and et al., “Pose transferrable person reidentification,” in CVPR, 2018, pp. 4099–4108.
17. Z. Zhong, L. Zheng, Z. Zheng, and et al., “Camera style adaptation for person re-identification,” in CVPR, 2018, pp. 5157–5166.
18. R. R. Varior, B. Shuai, J. Lu, D. Xu, and G. Wang, “A siamese long short-term memory architecture for human re-identification,” in ECCV, 2016, pp. 135–153.
19. W. Li, R. Zhao, T. Xiao, and X. Wang, “Deepreid: Deep filter pairing neural network for person re-identification,” in CVPR, 2014, pp. 152–159.
20. A. Hermans, L. Beyer, and B. Leibe, “In defense of the triplet loss for person re-identification,” arXiv preprint arXiv:1703.07737, 2017.
21. M. Ye, C. Liang, Z. Wang, Q. Leng, and J. Chen, “Ranking optimization for person re-identification via similarity and dissimilarity,” in ACM Multimedia (ACM MM), 2015, pp. 1239–1242.
22. M. Ye, C. Liang, Y. Yu, Z. Wang, Q. Leng, C. Xiao, J. Chen, and R. Hu, “Person reidentification via ranking aggregation of similarity pulling and dissimilarity pushing,” IEEE Transactions on Multimedia (TMM), vol. 18, no. 12, pp. 2553–2566, 2016.
23. A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In NIPS, 2012. 3.

24. K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In ICLR, 2015.3
25. C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In CVPR, 2015. 3.
26. K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In CVPR, 2016. 3, 7.
27. A. Wu, W. -S. Zheng, H. -X. Yu, S. Gong and J. Lai, "RGB-Infrared Cross-Modality Person Re-identification," 2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 5390-5399, doi: 10.1109/ICCV.2017.575.
28. Ye M, Lan X, Leng Q, et al. Cross-modality person reidentification via modality-aware collaborative ensemble learning[J]. IEEE Transactions on Image Processing, 2020, 29: 9387-9399.
29. Ye M, Lan X, Wang Z, et al. Bi-directional center-constrained top-ranking for visible thermal person re-identification[J]. IEEE Transactions on Information Forensics and Security, 2019, 15: 407-419.
30. Wang Z, Wang Z, Zheng Y, et al. Learning to reduce dual-level discrepancy for infrared-visible person re-identification[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019: 618-626.
31. Lu Y, Wu Y, Liu B, et al. Cross-modality person re-identification with shared-specific feature transfer[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 13379-13389.
32. Z. Wang, Z. Wang, Y. Zheng, Y.-Y. Chuang, and S. Satoh, “Learning to reduce dual-level discrepancy for infrared-visible person re-identification,” in CVPR, 2019, pp. 618–626.
33. Y.-C. Chen, W.-S. Zheng, J.-H. Lai, and P. Yuen. An asymmetric distance model for cross-view feature mapping in person re-identification. IEEE TCSVT, 2015. 4.

34. Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets[J]. Advances in neural information processing systems, 2014, 27.
35. Pingyang Dai, Rongrong Ji, Haibin Wang, Qiong Wu, and Yuyu Huang. Cross-modality person reidentification with generative adversarial training. In IJCAI, pages 677–683, 2018.
36. Guan'an Wang, Tianzhu Zhang, Jian Cheng, Si Liu, Yang Yang, and Zengguang Hou. Rgb-infrared crossmodality person re-identification via joint pixel and feature alignment. In Proceedings of the IEEE International Conference on Computer Vision, pages 3623–3632, 2019.
37. M. Ye, Z. Wang, X. Lan, and P. C. Yuen, "Visible thermal person reidentification via dual-constrained top-ranking," in International Joint Conference on Artificial Intelligence (IJCAI), 2018, pp. 1092–1099.
38. A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," arXiv preprint arXiv:1703.07737, 2017.
39. Alex D, Sami Z, Banerjee S, et al. Cluster loss for person re-identification[C]//proceedings of the 11th Indian conference on computer vision, graphics and image processing. 2018: 1-8.
40. LeCun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.
41. Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks[C]. International Conference on Neural Information Processing Systems. Curran Associates Inc. 2012:1097-1105.
42. Hinton G E, Srivastava N, Krizhevsky A, et al. Improving neural networks by preventing co-adaptation of feature detectors[J]. arXiv preprint arXiv:1207.0580, 2012.
43. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv:1409.1556, 2014.
44. Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions[C]. Proceedings of the IEEE conference on computer vision and pattern recognition. 2015:1-9.

45. Lin T Y, Dollár P, Girshick R, et al. Feature pyramid networks for object detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 2117-2125.
46. Kiefer J, Wolfowitz J. Stochastic Estimation of the Maximum of a Regression Function[J]. Annals of Mathematical Statistics, 1952, 23(3):462-466.
47. Kingma D P, Ba J. Adam: A method for stochastic optimization[J]. arXiv preprint arXiv:1412.6980, 2014.
48. Ruder S. An overview of gradient descent optimization algorithms[J]. arXiv preprint arXiv:1609.04747, 2016.
49. Chawla N V, Bowyer K W, Hall L O, et al. SMOTE: synthetic minority over-sampling technique[J]. Journal of Artificial Intelligence Research, 2002, 16(1):321-357.
50. Inoue H. Data Augmentation by Pairing Samples for Images Classification[J]. 2018.
51. Zhang H, Cisse M, Dauphin Y N, et al. mixup: Beyond Empirical Risk Minimization[J]. 2017.
52. Cubuk E D, Zoph B, Mane D, et al. AutoAugment: Learning Augmentation Policies from Data[J]. arXiv: Computer Vision and Pattern Recognition, 2018.

致 谢

攻硕期间参与项目、发表论文、参加测评及获奖情况

发表论文

1. 以第一作者撰写的论文《双端可共享网络的多模态行人重识别方法》并被《计算机工程与应用》中文核心期刊录用。

获奖情况

1.获得 2019-2020 年度硕士研究生二等奖学金

2.获得 2021-2022 年度硕士研究生二等奖学金