东北大学

硕士学位论文开题报告及论文工作计划书

学　　号：	1901758
姓　　名：	罗琪
导　　师：	焦明海
学科类别：	☑工学　□全日制专业学位
学科/工程领域:	计算机科学与技术
所属学院：	计算机科学与工程学院
研究方向：	多模态行人重识别
拟选题目：	行人重识别的双端共享网络多模态算法研究
选题时间：	2020　年　　月　　日

东北大学研究生院

年　　月　　日

填表说明

1、本表应在导师指导下如实填写。

2、学生在开题答辩前一周，将该材料交到所在学院、研究所。

3、按有关规定，没有完成开题报告的学生不能申请论文答辩。

4、全文正文均用小四号宋体，单倍行距，段前段后间距为0，如果页数不够，可以整页扩页，其他格式要求参见《东北大学硕、博士学位论文格式》。

一、前期工作基础（本节可以整页扩页）

课程学习及选题开题阶段，在导师指导下从事研究工作总结（不少于 2000 字）

研究生开学之初，焦明海老师为课题组的每一位同学分配了大的研究方向，让我们先去学习对应方向的基础知识，之后他再给予点评和把关。经过一段时间的学习和交流，我的研究方向确定为基于双端可共享网络的多模态行人重识别方法。明确了自己的研究方向以后，我就跟课题组的同学一起开始了有针对性的学习和研究。

1. 参加每周的研讨交流会

从研究生入学开始，在焦老师的组织下，我们就开始了每周一次的学术活动。在进行汇报之前，同学们都会介绍最近最新的与课题组相关的新概念、新技术、新方法，如卷积神经网络(CNN)、循环神经网络(RNN)、LSTM、GRU、SMOTE 不均衡算法等等，这很有助于拓宽我的知识面，以更开阔的视角去全面深入地理解所研究领域的理论、基础知识、核心技术等。在每周的学术讨论上，都会有一名同学讲解一篇比较新的论文，同学们都以 PPT 的形式进行汇报，言简意赅，深入浅出。课题组的成员比较多，大家你一言我一语，讨论的非常透彻。在每一次的组会上，老师都会给出许多高屋建瓴的点评。通过听取同学汇报和老师的点评，能够帮助我更好的了解当前国际的学术进展和最前沿的科学技术，为自己在以后的学习研究中积累了很多的理论性的知识。经过一次次的汇报、讨论和会后查阅相关资料，我尽快了解了行人重识别领域的基本研究方法和研究现状。聆听其他研究方向师兄师姐的毕业设计思路，扩展了我的思维和我的知识面，使我能够将其他领域的思想和知识应用同我的研究方向进行结合。让我产生了很多新颖的想法，对我研究的课题我也有了更加深刻的见解。同时，我和其他以行人重识别作为研究方向的同学组成了学习兴趣小组，在每周的学术研讨会之外，各个小组每两周进行一次学术交流会，对学习进展和所遇到的具体问题进行探讨。大家集思广益、充分讨论、发散思维，在交流中及时发现并解决问题，提出新的构想和观点。通过参加学术活动一方面，我认识到自己的很多不足，经过老师及师兄师姐们的指正，完善自己，让自己更懂得如何有效的学习和思考。在学术讨论之外，焦老师不仅教我们学习专业的只是，而且教我们怎样高效的学习以及做人的道理，真的很幸运遇到焦老师这样负责人的老师。

2.参加学术报告与讲座

截止目前我参加的学术讲座有：Dr.Nan Tang 教师的 Graph Steam Summarization、Amr EI Abbad 教授的区块链去神秘化，实现原子性跨链交易、加拿大皇后大学 Patrick Martin 教授的“Consumable Analytics for Big Data”、加拿大阿尔伯塔大学博士后李玉喜的“AlphaGo 核心技术及应用”、 英国 Ulster 大学 Dr.Zhiwei Lin 的“Multi-facet Consensus Measure of Rankings”、王国栋院士的“创新驱动发展，奔向光辉的 2049”、博士生会的“博士生 VR 技术与未来研讨会”、计算机学院的“CCF 走进高校,助力你的专业发展”、何万青教授的“高性能计算与 IT，互联网人才职业发展”、计算机学院的“计算机学院研究生会学术沙龙——开题那点事儿”、 清华大学林闯教授的“易经思维模型与网络数据计算模式”。特别需要指出的是，何万青教师的专题讲座对本人的启发很大，报告讲解了什么是高性能计算、互联网人才如何做好自己的职业规划，如何保持自己不被卷进技术更新的洪流中，并强调了科学知识和实际应用相互结合和促进的重要性。李玉喜博士的报告生动形象，深入浅出的讲解使人受益匪浅，讲出了 AlphaGo 核心的技术和原理，为我解开了机器战胜人类的本质原理。这一系列的学术讲座，有助于我了解毕业设计相关领域的前沿动态，拓宽眼界，让我能够以更发散的思维从事自己的学位论文研究工作。所谓万卷不离其宗，这些先进的技术和想法对我研究都有着很强的联系，在今后我研究的过程中是大有裨益。

3.阅读书籍文献资料

根据自己的研究方向，在焦老师的指导下阅读的书籍包括周志华的《机器学习》、李航的《统计学习方法》、Peter Harrington的《机器学习实战》与《数字图像处理的MATLAB实现》等；

阅读的学术文献包括:“Cross-Modality Person Re-Identification via Modality-aware Collaborative Ensemble Learning”、“Visible Thermal Person Re-Identification via Dual-Constrained Top-Ranking”、“AlignedReID++ Dynamically matching local information for person”、“Hi-CMD Hierarchical Cross-Modality Disentanglement for Visible-Infrared”、“RGB-IR Cross-modality Person ReID based on”、“Multi-modal deep feature learning for RGB-D object detection”、“Dynamic Dual-Attentive Aggregation Learning”、“Deep Learning for Person Re-identification”和“Learning to Reduce Dual Level Discrepancy for Infrared Visible Person Re Identification”等20余篇。

《统计学习方法》是李航的一本书，是比较基础经典的一本书，书中更多的是对基础传统机器学习的理论介绍，没有任何代码，这算是对代码的补充。另外一本书《机器学习实战》，是包含部分理论与代码的书，也很经典，我看了下代码，因为这两本书都是基础类的书，重叠部分还是很多的。但是整体来说《统计学习方法》理论东西更多，再配合代码就很完美了。我一直认为，机器学习必须理论与代码都要掌握，学会理论才算真的入门，会代码的才会实际应用。光会代码不会理论就只能调用 API 了，碰到实际问题多半也解决不了，而且根本无法理解算法的美。“Deep Learning for Person Re-identification:A Survey and Outlook”是一篇关于行人重识别的综述文章，通过行人重识别领域的研究，将该领域分为封闭世界（closed-world）和开放世界（open-world）。从三个不同角度对封闭世界进行深入分析，分别为深度特征表示学习，深度度量学习和排名优化，同时针对当前经常用到的评价标准的不足，提出了一种新的度量标准 mINP 用于测量罚分以找到最难的正确匹配项。最后，文章作者提出了一种对于行人重识别领域的 baseline，对于多数行人重识别问题有较高的准确率。“RGB-Infrared Cross-Modality Person Re-Identification” 本文评估了现有的跨域模型，包括三种常用的神经网络结构（一流、两流和非对称 FC 层），并分析了它们之间的关系。作者进一步提出了深度零填充的方法来训练一个单流网络，使其能够自动改进网络中特定于域的节点，从而实现跨模态匹配。同时文章作者首次提出了支持 RGB-IR 交叉模态 Re-ID 的标准基准 SUSU-MM01 数据集，包括来自 6 个摄像头的 491 个 identities 的 RGB 和 IR 图像，共给出 287628 个 RGB 图像和 15792 个红外图像，并进行了广泛的实验，如今该数据集成为跨模态行人重识别领域应用最广泛的数据集。

通过阅读相关文献和书籍，大致了解了研究方向上面临的挑战和亟待解决的问题，如跨模态的差异问题，特征融合问题以及分类器训练问题等等。明确设计方向，为后续研究基于多模态行人重识别的研究提供了理论基础和科学依据。

4. 毕设开题辅导

进入毕业设计开题阶段，为了使同学们明确自己的研究方向，焦老师每周都会组织我们开会，汇报内容包括：毕业设计拟定的题目，出发点与整体研究思路，老师会对每位同学进行指导，提出意见与建议。通过多次汇报，我们受益匪浅，不断完善我们毕业设计的整体框架，为下一步的工作奠定良好的基础。

二、选题依据（本节可以整页扩页）

课题背景、选题依据、课题研究目的、理论意义和应用价值（工学硕士）/工程背景和实用价值（专业学位硕士）（不少于 1000 字）

1. 课题背景和选题依据

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借助于深度神经网络，取得了一定的效果，但大多数网络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，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。第一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。第二类方法使用两个不同的网络提取特征后，需要处理不同模态的特征，如何有效的减小模态差异成为了研究的重点和难点。

通过对当前多模态行人重识别的研究，对双端可共享网络做了一些改进，使得重识别网络可以更有效地处理跨模态信息。提出的双端可共享行人重识别网络有以下几点优势:

（1）针对真实数据中行人姿态不对齐问题，提出一种针对训练数据的数据预处理方法，使得数据更接近真实场景下的图像。将处理后的数据与原数据一起作为训练数据作为输入，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关注块（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，提高计算效率。传统的卷积神经网络（CNN）是在局部空间做局部操作，大范围的提取特征需要依靠重复堆叠，效率太低，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。当前的行人重识别网络大多采用三元组损失函数来训练网络，在数据集不是很大时，三元组损失可以有很好的效果，因为三元组损失每次选择一个与锚样本同类不同模态的样本和同模态不同类的样本这两个样本可以在数据集比较小的时候对整个类起到一定的作用，但数据量很大时，需要选择的三元组数量会增长，使得训练变得复杂。本方法中使用的聚类损失并不从一批图像中提取所有图像的累积贡献，而是提取对损失贡献最大的样本，从而使训练批次中的难样本对损失函数有直接贡献。

2. 课题研究目的

对于全天候监控的监控系统，传统的单模态行人重识别网络无法适用于红外图像的行人重识别，为了能够在任意模态的数据集中实现行人重识别，需要设计一种网络结构能够同时处理红外图像和可见光图像，并且能够有效的识别目标行人。当前对于跨模态行人重识别的研究中，使用对抗生成网络生成相对应的模态，然后使用单网络进行训练，虽然该方法有一定的准确率，但由于涉及到对抗生成网络的训练，使得训练成本增加，不利于模型的部署。另一种方法使用双端网络输入两种模态的图片，然后进行特征融合，进入到共享网络进行训练，这种方法由于后半部分的网络参数是共享的，不仅要能够处理不同的模态信息，也要有

一定的准确率，因此如何优化网络结构以及如何选择更有效的损失函数成为该领域的重点和难点，本论文会基于双端可共享网络进行网络结构的优化和损失函数的改进，使得训练后的网络在行人重识别任务上有更高的准确率。

3. 理论意义

监控图像中，可见光图像和红外图像由于摄取的光的波长不同，导致两类图像之间存在模态差异。使用单模态行人重识别网络无法有效的识别红外图像的行人，因此需要一种双端网络来分别输入两种模态的图像，并经过特征提取和特征融合后进行分类。该网络结构的难点在于如何有效的减小或消除模态差异，同时有效的辨别类间差异，从而更好的实现图像的分类。同时由于监控系统的普及，监控图像的数据量越来越大，传统的三元组损失如果面对大规模的数据量，需要选择的三元组数量也随之增大，使得训练变得复杂。使用聚类损失可以使每个样本都可以对损失函数具有贡献，能够优化网络的训练过程。

对于上述问题的研究可以有效的改善多模态行人重识别的效率以及准确率，对于多模态特征的融合以及损失函数的设计有重要的参考价值。

4. **应用价值**

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。多模态行人重识别的研究正是基于以上的现实问题而提出的，而双端可共享网络对于解决这一问题起到了很好的效果，通过对该网络的改进可以提高网络在多模态数据集中的准确率，从而在海量数据前更高效的实现目标人物的搜寻。

三、文献综述（本节可以整页扩页）

国内外研究现状、发展动态描述（不少于 1000 字）；所阅文献的查阅范围及手段，附参考文献（不少于 10 篇，其中近 3 年文献不少于 5 篇，英文文献不少于 3 篇，全部按照标准格式列出，并在文中顺序标注）

1. 国内外研究现状、发展动态描述

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。

文献[1]提出一种域选择的子网络，可以自动选择样本所对应的模态，同时提出了包含 RGB 图像和红外图像的数据库 SYSU-MM01。作者假设存在一种域选择子网络可以自动选择相对应的样本作为输入，使得所有结构都可以用单流结构来表示。该方法将 RGB 图像和红外图像作为两个不同域的输入，使用深度零填充后放入上述网络中，使得所有输入都可以用单流结构来表示。

文献[2]利用 Classification/Identification loss 和 verification loss 来训练网络，其网络示意图如下图所示。网络输入为若干对行人图片，包括分类子网络(Classification Subnet)和验证子网络(Verification Subnet)。分类子网络对图片进行 ID 预测，根据预测的 ID 来计算分类误差损失。验证子网络融合两张图片的特征，判断这两张图片是否属于同一个行人，该子网络实质上等于一个二分类网络。经过足够数据的训练，再次输入一张测试图片，网络将自动提取出一个特征，这个特征用于行人重识别任务。

文献[3]为了解决图像不对齐问题，先用姿态估计的模型估计出行人的关键点，然后用仿射变换使得相同的关键点对齐。一个行人通常被分为 14 个关键点，这 14 个关键点把人体结果分为若干个区域。为了提取不同尺度上的局部特征，作者设定了三个不同的 PoseBox 组合。之后这三个 PoseBox 矫正后的图片和原始为矫正的图片一起送到网络里去提取特征，这个特征包含了全局信息和局部信息。

文献[4]提出了一种全局-局部对齐特征描述子(Global-Local-Alignment Descriptor, GLAD)，来解决行人姿态变化的问题。GLAD 利用提取的人体关键点把图片分为头部、上身和下身三个部分。之后将整图和三个局部图片一起输入到一个参数共享 CNN 网络中，最后提取的特征融合了全局和局部的特征。为了适应不同分辨率大小的图片输入，网络利用全局平均池化(Global average pooling, GAP)来提取各自的特征。

文献[5]中使用的 Spindle Net 也利用了 14 个人体关键点来提取局部特征。Spindle Net 直接利用这些关键点来抠出感兴趣区域(Region of interest, ROI)。首先通过骨架关键点提取的网络提取 14 个人体关键点，之后利用这些关键点提取 7 个人体结构 ROI。网络中所有提取特征的 CNN 参数都是共享的，这个 CNN 分成了线性的三个子网络 FEN-C1、FEN-C2、FEN-C3。对于输入的一张行人图片，有一个预训练好的骨架关键点提取 CNN 来获得 14 个人体关键点，从而得到 7 个 ROI 区域，其中包括三个大区域（头、上身、下身）和四个四肢小区域。这 7 个 ROI 区域和原始图片进入同一个 CNN 网络提取特征。原始图片经过完整的 CNN 得到一个全局特征。三个大区域经过 FEN-C2 和 FEN-C3 子网络得到三个局部特征。四个四肢区域经过 FEN-C3 子网络得到四个局部特征。之后这 8 个特征按照图示的方式在不同的尺度进行联结，最终得到一个融合全局特征和多个尺度局部特征的行人重识别特征。

以上的局部特征对齐方法都需要一个额外的骨架关键点或者姿态估计的模型。而训练一个可以达到实用程度的模型需要收集足够多的训练数据，这个代价是非常大的。为了解决以上问题，文献[6]提出一种基于 SP 距离的自动对齐模型（AlignedReID），在不需要额外信息的情况下来自动对齐局部特征。该方法将输入图片提取的特征与锚图片的提取特征进行划分，并对每个部分计算与另一图片的特征各部分的距离，而采用的方法就是动态对齐算法，或者

也叫最短路径距离。这个最短距离就是自动计算出的 local distance。

文献[7]提出一种基于跨层特征连接和双模态三重损失的视热交叉模态人物识别模型。模型的主干网络为 ResNet50，并采用双流结构，主干部分分别输入 RGB 图像和热力图像，然后以参数共享的方式将两流的顶层全连接层嵌入到共同的特征空间中，同时将 CNN 模型中间层的中层特征与最终的骨干特征融合，提高人特征的识别力。最后两个模块都采用双模态三重损失(Ld_tri)和分类损失(Lsoftmax)进行网络训练。该方法可以同时解决跨模态差异和模内差异。

文献[8]提出一种动态双注意聚合学习网络（DDAG），DDAG 是在双流网络上开发的，包含了用于分离部分聚合特征学习的模内加权部分和用于共享全局特征学习的跨模图结构部分。双流网络中的特征提取器是参数共享的。为了处理局部特征和全局特征，分别使用模态内加权聚合（IWPA）和跨模态图结构注意（CGSA）来进行网络的训练。同时进一步引入了一种无参数的双聚合学习策略来自适应地聚合两个组件。

文献[9]提出一种分层交叉解析方法（Hi-CMD），目的是排除姿态、光照这些冗余特征（ID-excluded）的影响，提取出更加有判别力的体态、衣着等信息。 该方法包含了两个核心模块：ID-PIG 网络和 HFL 模块。ID-preserving Person Image Generation（ID-PIG）网络，在保证行人 ID 不变的条件下，改变它的姿态、光照属性。Hierarchical Feature learning（HFL）模块用于确保编码器能够提取具有判别力的特征，对姿态、光照变化具有鲁棒性。

文献[10]提出一种双流网络结构，该结构使用两个基于 Resnet50 的特征提取器分别提取两个模态的特征，将两个模态的特征提取后经过全连接层，该全连接层为两个网络所共享的，目的是进行特征融合。在进行训练使采用双向训练策略来约束整个学习过程。双向排序损失考虑了两种关系:可见光-红外关系(一个锚可见图像，两个红外图像)和红外-可见光关系(一个锚红外图像，两个可见图像)。为了解决跨模态变化引起的类内距离大于类间距离的问题，作者使用难样本挖掘来提高识别能力。基本的思想是比较一个正的可视-红外对的距离和所有相关的负可视-红外对的最小距离，这样可以保证模态引起的类内距离小于同模态的类间距离。

文献[11]提出一种模态感知协作的中层可共享的双端网络，将 Resnet50 的第一层卷积层作为各自模态的浅层特征提取器，后四层卷积层作为共享网络，输入融合两个模态的浅层特征后继续进行特征提取。在进行分类时，使用两个单模态分类器来辅助多模态分类器的学习，使用集成学习损失和一致性损失来训练分类器。并使用三元组损失训练网络，该方法大大降低了训练难度。

文献[12]提出一种新颖的端到端对齐生成网络(AlignGAN)，方法可以概括为使用循环 GAN 网络生成与真实 RGB 图像对应的虚拟 IR 图像，再将假的 IR 图像与真实的 IR 图像做验证。模型由像素生成器、特征生成器和联合鉴别器，三要素之间进行大小博弈。缓解跨模态和模态内的差异，能够学习到身份一致性。

文献[13]提出一种新型的交叉模式生成对抗网络（cmGAN）来处理跨模态信息。针对识别信息不足的问题，设计了一种基于前沿生成对抗训练的识别器，从不同的模式中学习识别特征表示。为了解决大规模的跨模态度量学习问题，将识别损失和跨模态三重损失结合起来，在最大化实例间的跨模态相似性的同时，将类间模糊性最小化。使用标准的深度神经网络框架，可以对整个 cmGAN 进行端到端的训练。

现有的跨模态行人重识别技术在公共数据集上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，对于理论研究和实际应用价值非常有意义。

2. 查阅文献范围及手段

IEEE 数据库	2008-2020
ACM 数据库	2008-2020
Google 学术	1980-2020
Springer 数据库	2005-2020

Ei Village	2009-2020
SCIE（Web of Science）	2005-2020
Elsevier 数据库	2009-2020
中国优秀博硕士论文数据库	2009-2020
中国科技期刊数据库（维普）	2009-2020
中国学术期刊全文数据库	2009-2020
东北大学图书馆自然科学阅览室图书	

3. 参考文献

[1] Wu A, Zheng W S, Yu H X, et al. Rgb-infrared cross-modality person re-identification[C]//Proceedings of the IEEE international conference on computer vision. 2017: 5380-5389.

[2] Mengyue Geng, Yaowei Wang, Tao Xiang, Yonghong Tian. Deep transfer learning for person reidentification[J]. arXiv preprint arXiv:1611.05244, 2016.

[3] Liang Zheng, Yujia Huang, Huchuan Lu, Yi Yang. Pose invariant embedding for deep person reidentification[J]. arXiv preprint arXiv:1701.07732, 2017.

[4] Longhui Wei, Shiliang Zhang, Hantao Yao, Wen Gao, Qi Tian. Glad: Global-local-alignment descriptor for pedestrian retrieval[J]. arXiv preprint arXiv:1709.04329, 2017.

[5] Haiyu Zhao, Maoqing Tian, Shuyang Sun, Jing Shao, Junjie Yan, Shuai Yi, Xiaogang Wang, Xiaoou Tang. Spindle net: Person re-identification with human body region guided feature decomposition and fusion[C]. CVPR, 2017

[6] Zhang, X., Luo, H., Fan, X., Xiang, W., Sun, Y., Xiao, Q., ... & Sun, J. (2017). AlignedReID: Surpassing Human-Level Performance in Person Re-Identification. arXiv preprint arXiv:1711.08184.

[7] Liu H, Cheng J, Wang W, et al. Enhancing the discriminative feature learning for visible-thermal cross-modality person re-identification[J]. Neurocomputing, 2020.

[8] Ye Mang, Shen Jianbing, Crandall David J, Shao Ling, Luo Jiebo Dynamic. Dual-Attentive Aggregation Learning for Visible-Infrared Person Re-Identification[J]. European Conference on Computer Vision (ECCV),arXiv:2007.09314.

[9] S. Choi, S. Lee, Y. Kim, T. Kim and C. Kim, "Hi-CMD: Hierarchical Cross-Modality Disentanglement for Visible-Infrared Person Re-Identification," 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 2020, pp. 10254-10263, doi: 10.1109/CVPR42600.2020.01027.

[10] M. Ye, X. Lan, Z. Wang and P. C. Yuen, "Bi-Directional Center-Constrained Top-Ranking for Visible Thermal Person Re-Identification," in *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 407-419, 2020, doi: 10.1109/TIFS.2019.2921454.

[11] M. Ye, X. Lan, Q. Leng and J. Shen, "Cross-Modality Person Re-Identification via Modality-Aware Collaborative Ensemble Learning," in IEEE Transactions on Image Processing, vol. 29, pp. 9387-9399, 2020, doi: 10.1109/TIP.2020.2998275.

[12] G. Wang, T. Zhang, J. Cheng, S. Liu, Y. Yang and Z. Hou, "RGB-Infrared Cross-Modality Person Re-Identification via Joint Pixel and Feature Alignment," 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea (South), 2019, pp. 3622-3631, doi: 10.1109/ICCV.2019.00372.

[13] DAI, Pingyang, et al. Cross-modality person re-identification with generative adversarial training. In: IJCAI. 2018. p. 2.

四、研究内容（本节可以整页扩页）

1．研究构想与思路、主要研究内容及拟解决的关键问题（不少于 1000 字）

1.1 研究构想与思路

随着监控系统的普及，监控摄像头将越来越多的监控图像传入到数据库，这些图像不仅包括可见光图像，同时也包括红外图像。在进行行人重识别时，不仅由于数据量大所带来的检测过程复杂度高，而且还由于图像之间存在着模态差异而导致的识别率下降，如何确保行人重识别网络在不降低识别准确率的同时，还可以处理由模态不同而导致的模态差异，成为当下行人重识别领域的重点。

当前的多模态行人重识别算法可以大致分为两类，一类是借助于 GAN 网络进行图片生成来辅助网络实现重识别，另一类是使用共享网络提取共有特征后对特征进行融合，再进行识别。第一类的算法中，比如 AlignGAN、cmGAN 等，虽然有一定的准确率，但由于涉及到 GAN 网络的训练，使得训练复杂度提高，不利于模型的推广。第二类算法的主要思路是先分别提取两种模态图片的特征，然后对特征进行特征融合，再输入到共享网络中学习，该类方法虽然训练复杂度较低，但在准确率提升以及网络结构优化方面还可以有大量改善的空间，因此，本文通过对当前行人重识别领域主流算法的研究，分析了各个方法的优点和不足，提出一种双端可共享的多模态行人重识别方法。

本文的研究构想分为三部分，分别是：

（1）由于行人重识别领域的公共数据集较少，现有的数据集无法涵盖真实场景的各种图像，比如姿态不对齐、比例不一致等，若不对数据集进行数据增强会使得训练所得网络没有鲁棒性，在实际场景中效果不佳。因此可以尝试对数据进行处理，使得处理后的图像更接近实际的场景。

（2）在进行特征融合后，共享网络中的特征提取器会挖掘输入的共有特征，普通的卷积操作一般都是在附近区域内执行，对于融合后的特征，两个相关联的样本点很可能会在空间上有一定的距离，使用一般的卷积层很难准确的挖掘到这样的两个样本点之间的关系。为了捕获长范围依赖，本文在网络中嵌入非局部注意力网络来解决这一问题，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，提高计算效率非局部注意力块网络结构如图 2 所示。使得网络的训练效率更高。

（3）一般的行人重识别网络在进行距离度量是使用三元组损失，三元组损失通过挖掘难样本来指导网络的训练，在数据量不大的情况下，该方法有很好的效果，但随着数据量的增长，选择出的三元组数量会更多，使得训练复杂度提高。随着训练的进行，网络更多的关注那些难样本，而忽略大部分的普通样本。因此，本文使用聚类损失来替换三元组损失，该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，从而提高训练效率。

整体的网络结构如图 3 所示。

1.2 主要研究内容

根据研究思路和构想，本文确定了以下的研究内容：

（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预处理方法，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）为了提取特征图中的长范围依赖，在 Resnet50 共享网络的后四个卷积层中嵌入非局部关注块（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，提高计算效率。传统的卷积神经

网络（CNN）是在局部空间做局部操作，大范围的提取特征需要依靠重复堆叠，效率太低，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。由于在数据量很大时，三元组损失的难样本挖掘需要选择的三元组数量会增长，使得训练变得复杂，训练效果不理想。本方法中使用的聚类损失基于均值来计算距离，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所有样本对损失函数都有间接贡献。

1.3 拟解决的关键问题

（1）如何对数据进行预处理，使得数据更接近真实场景的数据，同时使得使用预处理数据训练的模型更具有鲁棒性。

（2）如何在特征提取的过程中扩大感受野，汇聚更多有用的信息，突出重点区域，同时能够消除噪声，提高特征提取的效率。

（3）如何有效的减小或消除输入图片之间的模态差异，同时提高行人重识别的准确率。

（4）如何更高效的训练多模态双端网络，尽管三元组损失可以有效的训练网络，但在大规模数据集上三元组损失的训练效率并不好，应该使用其它更高效的损失函数代替三元组损失来对网络进行训练。

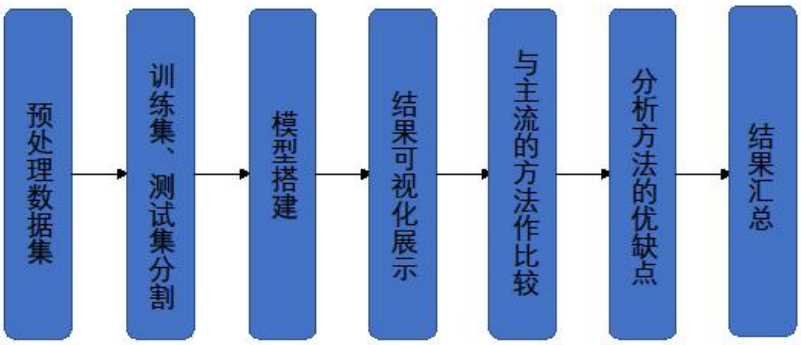

图 1 研究思路

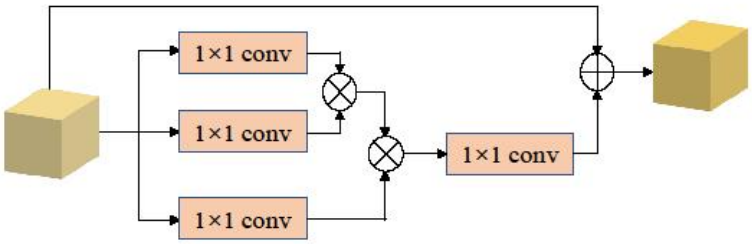

图 2 非局部注意力块网络结构

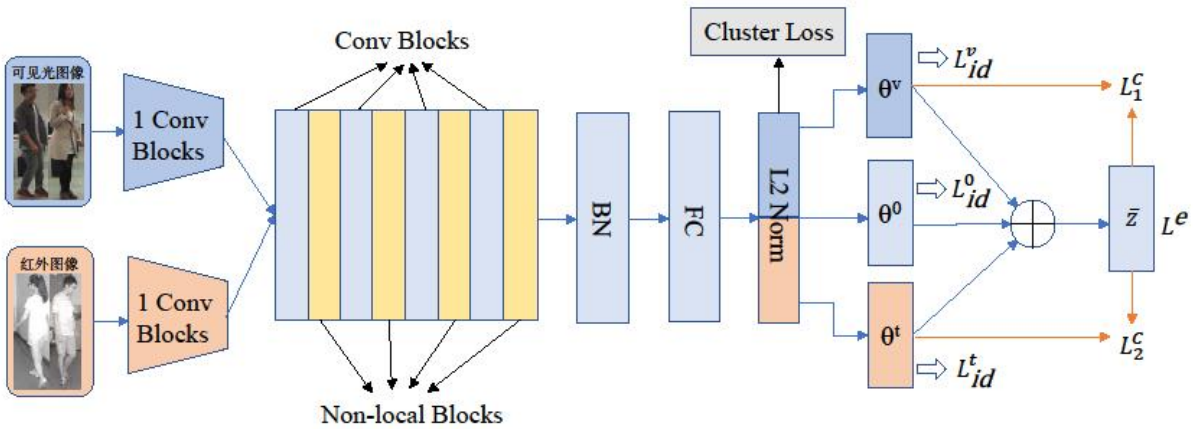

图 3 网络结构

2．拟采取的研究方法、技术路线、实施方案及可行性分析

2.1 拟采取的研究方法与技术路线

（1）一般的数据增强方法包括随机翻转、随机遮挡、裁剪等，本文类比常用的数据预处理方法，针对行人重识别中遇到的具体问题，尝试对数据集的图像进行缩减和填充，以模拟真实场景下的图像，将处理后的图像与数据集混合后，输入网络进行训练；

（2）本文借助于深度学习领域中捕获长范围依赖的问题，在特征提取网络中嵌入非局部注意力块，用于捕获长范围依赖。非局部注意力块在计算机视觉和自然语言处理方面被证明是有效的，在共享网络中嵌入非局部注意力块会使特征提取更有效率；

（3）本文使用基于特征均值的聚类损失训练网络，聚类损失在机器学习领域被广泛使用，对于相似样本的聚类可以发挥重要作用，使用聚类损失训练网络，使得每个样本都能对网络训练有间接的贡献；

在本文采用的基于特征均值的聚类损失函数中，设 $f^v(x)$ 、$f^t(x)$ 分别表示输入 x 经过网络后得到的可见光图像特征和红外光图像特征，对于同一个批次中 K 个相同模态的样本，某一身份 i 不同模态的平均特征可分别表示为：

$$f_i^{vm}=\frac{\sum_K f^v(x)}{K} \quad 、 \quad f_i^{tm}=\frac{\sum_K f^t(x)}{K} \quad ;$$

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，同时为了挖掘难样本，因此只选择距离最大值：

$$d_i^{intra(v-v)}=\max_K \| f^v(x)-f_i^{vm}(x) \|_2^2 \quad 、 \quad d_i^{intra(t-t)}=\max_K \| f^t(x)-f_i^{tm}(x) \|_2^2$$

考虑到跨模态的同类样本，因此有：

$$d_i^{intra(v-t)}=\max_K \| f^v(x)-f_i^{tm}(x) \|_2^2 \quad 、 \quad d_i^{intra(t-v)}=\max_K \| f^t(x)-f_i^{vm}(x) \|_2^2$$

同样地，一个身份的类间距离由该身份的特征均值到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：

$$d_i^{inter(v-v)}=\min_{\forall id\in P, id\neq i} \| f_i^{vm}(x)-f_{id}{}^{vm}(x) \|_2^2 \quad 、 \quad d_i^{inter(t-t)}=\min_{\forall id\in P, id\neq i} \| f_i^{tm}(x)-f_{id}{}^{tm}(x) \|_2^2$$

类似地，跨模态下的类间距离可表示为：

$$d_i^{inter(v-t)}=\min_{\forall id\in P, id\neq i} \| f_i^{vm}(x)-f_{id}{}^{tm}(x) \|_2^2 \quad 、 \quad d_i^{inter(t-v)}=\min_{\forall id\in P, id\neq i} \| f_i^{tm}(x)-f_{id}{}^{vm}(x) \|_2^2$$

由以上两类距离可以得到四个聚类损失函数：

$$L_C^{v-v}=\sum_{\mathrm{i}}^{P}\max((d_i^{intra(v-v)}-d_i^{inter(v-v)}+\alpha),0)$$

依此类推，可求得 L_C^{t-t} 、L_C^{v-t} 和 L_C^{t-v} 。最终得到总的聚类损失为：

$$L_C=L_C^{v-v}+L_C^{t-t}+L_C^{v-t}+L_C^{t-v} /$$

本文的技术路线如图 4 所示。主要分为三部分，第一部分是对数据集中图像的处理，将图像进行缩小和填充，使其能模拟真实场景下的数据。第二部分是对网络结构进行的改进，在特征提取器嵌入非局部注意力块，使网络能够有效地获取特征图的长范围依赖。第三部分是对损失函数的改进。使用基于特征均值距离的聚类损失函数，可以有效的解决三元组损失函数的不足，使得每个样本都对网络训练有贡献，提高训练效率。

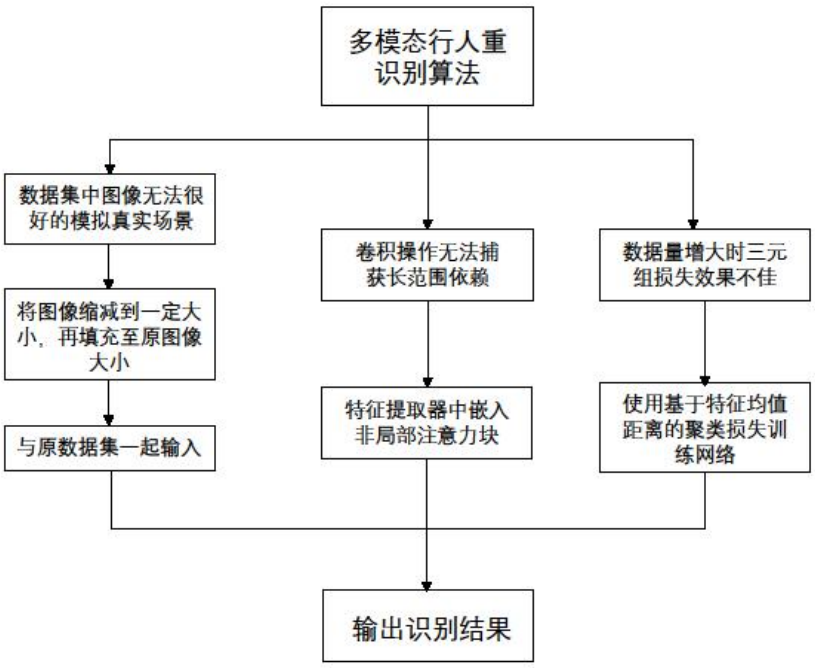

图 4 技术路线图

2.2 实施方案

为了解决跨模态行人重识别存在的问题，本文拟通过三个方面来实施，分别是数据处理、网络结构和训练方法。

（1）对数据集进行预处理，使得数据集中的图像接近真实场景的图像，而不是只有相同大小比例的数据，这样使得训练出的算法更具有鲁棒性。

（2）针对特征提取网络无法扩大感受野，使用非局部注意块，这样可以在提取特征时增加感受野，使得网络更多的关注于突出部分，同时可以将大范围内的与当前样本点有关联的点联系起来，这样使特征提取时效率更高。

（3）将三元组损失改为聚类损失，这样可以适应大数据集的训练。聚类损失中能够将一个批次中的数据都对损失函数有所贡献，使得算法收敛更快，训练效率更高。

2.3 可行性分析

（1）在计算机视觉领域，训练模型之前都会进行数据预处理，通过处理后的数据会使得训练出的模型更具有鲁棒性。一般的数据增广方法包括随机擦除、随机翻转、增缩、分割、填充等。本文提出的数据预处理方法主要使用缩减和填充，经过处理后使得数据与真实场景下的图像接近，从而使模型的鲁棒性提高。

（2）非局部注意力块（Non-local Attention Block）是为了解决卷积和循环算子在空间和时间上局部操作的不足而提出的。相比较于不断堆叠卷积和 RNN 算子，非局部操作直接计算两个位置(可以是时间位置、空间位置和时空位置)之间的关系即可快速捕获长范围依赖，这种计算方法是一种泛化的自相关矩阵，同时非局部操作计算效率很高，要达到同等效果，只需要更少的堆叠层，而且非局部操作可以保证输入尺度和输出尺度不变，这种设计可以很容易嵌入到目前的网络架构中。越来越多的实验表明，非局部注意力块可以有效的增加感受野，将注意力集中在感兴趣的地方，突出重点区域，这样的特性适应于行人重识别中提取行人特征的过程。

（3）三元组损失随着数据集变得更大，需要的三元组数量会立方增长，使得训练成本增加。聚类损失是根据 K 均值聚类和线性判别分析而提出的，其目的是最小化类内差异，最大化类间差异。三元组损失仅将函数建立在少数三个样本上，而本文使用的聚类损失函数即使

只考虑难样本，但由于它们的距离是基于平均值计算的，因此每个样本都对损失函数有间接的贡献，从而使得整个集群内部距离更小，集群间距离更大。

五、预期研究成果（本节可以整页扩页）

对所研究的成果进行阐述，同时要对与前文研究内容的相关性及与前人（他人）研究成果的差异性进行描述

1 由于真实场景下会因为摄像头的角度、距离等因素导致人物图像比例不一致，导致识别准确率下降，本文提出的数据预处理方法可以更接近于真实场景的图像。一般的数据预处理方式是对图像进行随机擦除、随机翻转、遮挡以及填充，对于普通的计算机视觉任务的训练有一定效果，行人重识别任务使用普通的数据增强方法已不能很好的模拟真实数据，所以将数据中人物的比例进行调整并进行填充，会很好的模拟真实数据，从而提高网络的鲁棒性。

2 普通的卷积网络对于特征提取启到至关重要的作用，但大多数卷积层都是依靠比较小的卷积核来执行，使得当前样本点只能和该样本点周围的点关联起来，想要捕获长范围的依赖，只能依靠重复的堆叠，这样效率很低。非局部注意力块的提出就是为了解决这样的问题。本文中引入非局部注意力块，可以很好的将一个样本点在全局范围内与它所有有关联的样本点联系起来，快速捕获长范围依赖，计算效率更高。

3 行人重识别中主要使用三元组损失函数来约束网络训练，三元组损失只将函数建立在少数三个样本之上，虽然进行了难样本挖掘，但对于数据量很大的情况，三元组损失无法综合大部分样本对于模型的贡献。使用聚类损失可以很好的解决这一问题，虽然聚类损失也只考虑了难样本，但由于在进行距离计算时是基于整个批次的平均值来计算的，因此间接的汇集了所有样本的贡献，这使得网络训练时能够更快地收敛。

六、研究条件（本节不允许扩页）

1．所需实验手段、研究条件和实验条件

以实验室为依托，以深度学习服务器支撑，查阅相关资料，开展针对基于双端共享网络的多模态行人重识别研究。依据现有的条件，参照之前的相关的论文和代码，使用 Ubuntu 下搭建的基于 Pytorch 模块，通过实际的数据输入，实现本系统基本功能，然后通过对代码的修改实现本文提出的要加入的元素。同时将搜集的历史流量数据导入到 Pycharm 软件中进行处理。在实现时，选择经典的双端网络和衡量指标与本文所设计的算法进行对比，并给出性能评价结果。

PyTorch 是一个开源的 Python 机器学习库，基于 Torch，用于自然语言处理等应用程序。2017 年 1 月，由 Facebook 人工智能研究院（FAIR）基于 Torch 推出了 PyTorch。它是一个基于 Python 的可续计算包，具有强大的 GPU 加速的张量计算（如 NumPy），同时包含自动求导系统的深度神经网络。PyTorch 的前身是 Torch，其底层和 Torch 框架一样，但是使用 Python 重新写了很多内容，不仅更加灵活，支持动态图，而且提供了 Python 接口。它是由 Torch7 团队开发，是一个以 Python 优先的深度学习框架，不仅能够实现强大的 GPU 加速，同时还支持动态神经网络，这是很多主流深度学习框架比如 Tensorflow 等都不支持的。PyTorch 既可以看作加入了 GPU 支持的 numpy，同时也可以看成一个拥有自动求导功能的强大的深度神经网络。

PyCharm 是一种 Python IDE，带有一整套可以帮助用户在使用 Python 语言开发时提高其效率的工具，比如调试、语法高亮、Project 管理、代码跳转、智能提示、自动完成、单元测试、版本控制。此外，该 IDE 提供了一些高级功能，以用于支持 Django 框架下的专业 Web 开发。

2．所需经费，包含经费来源、开支预算（工程设备、材料须填写名称、规格、数量）

书籍	200 元
打印、复印相关资料	500 元
总计	700 元

七、工作计划（本节不允许扩页）

序号	阶段及内容	工作量估计（时数）		起止日期	阶段研究成果
1	调研（阅读行人重识别的相关文献）	500		2020.10-2020.12	把握研究方向完成文献综述
2	需求分析与问题定义（对双端可共享网络模型和行人重识别机制进行分析）	400		2021.01-2021.03	完成分析过程提出初步方案
3	总体设计（对方案的整体概要进行设计）	400		2021.04-2021.06	依据模型提出算法概要设计
4	详细设计（对网络模型和行人重识别机制进行详细设计）	1000		2021.07-2021.09	完成网络模型和算法的详细设计
5	实现阶段（熟悉原型平台，搭建原型环境。）	1000		2021.10-2021.12	实现基于双端网络的主体架构
6	测试与性能评价（测试算法的性能，依据结果对压模型算法进行修改和完善）	200		2022.01-2022.03	修改并完善算法，进行性能评价
7	总结、论文写作（论文撰写阶段，进行论文成篇工作）	200		2022.04-2022.06	对整体工作进行总结、完善，完成毕业论文
		合计	4300		