浙江工业大学

硕士学位论文

监控视角下基于深度学习的人物识别方法研究

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

提交日期：2020 年 06 月

Deep Learning Based Person Recognition Methods under Surveillance Cameras

Dissertation Submitted to

Zhejiang University of Technology

in partial fulfillment of the requirement

for the degree of

Master of Engineering

by

Hao LI

Dissertation Supervisor: Prof. Yu-bo ZHAO

Jun., 2020

浙江工业大学学位论文原创性声明

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

作者签名：李灏　　日期：2020年5月

学位论文版权使用授权书

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

本学位论文属于 1、保密□，在一年解密后适用本授权书。

2、保密□，在二年解密后适用本授权书。

3、保密□，在三年解密后适用本授权书。

4、不保密☑。

（请在以上相应方框内打“√”）

作者签名：李灏　　日期：2020年 5 月

导师签名：　　日期：2020年 5 月

中图分类号 TH391　　学校代码 10337

UDC 004.8　　密级 公开

研究生类别 全日制学术型型硕士研究生

浙江工业大学

工学硕士学位论文

监控视角下基于深度学习的人物识别方法研究

Deep Learning Based Person Recognition Methods under Surveillance Cameras

作者姓名 李灏　　第一导师 赵云波 教 授

学位类型 工学硕士

学科专业 控制科学与工程　　培养单位 信息工程学院

研究方向 计算机视觉

答辩日期： 2020 年 5 月 17 日

监控视角下基于深度学习的人物识别方法研究

摘　要

监控视角下人物识别是智能安防领域内重点研究方向，高效准确的人物识别算法能在公安刑侦等需要追踪特定目标的场景中发挥重要作用。然而监控摄像头拍摄的图片带有复杂的背景信息及噪声干扰，造成传统人物识别方法无法有效识别。深度学习方法相比传统人物识别方法在行人目标检测和识别领域在准确性以及鲁棒性上存在很大提升。因此，基于深度学习的人物识别方法引发大量关注。

本文从卷积网络结构，排序损失以及身份损失等三个方面对单模态行人重识别以及跨模态行人重识别的人物识别算法进行改进。改进方法在行人重识别数据集实验中相比原有方法在精度上均有提升。同时本文基于深度学习方法设计一套监控视角下人物识别系统，为智能安防项目提供一套可行方案。本文主要工作如下：

(1) 提出了基于表征学习的中层特征扩展方法，缓解了单模态行人重识别中卷积网络模型对训练数据集过拟合问题。本文对卷积网络模型进行改造，在骨干网络基础上抽取中层特征进行分组池化操作后，将其与高层特征拼接输出。实验证明所提方法可以有效提升模型泛化能力，提高人物识别精度。

(2) 提出了基于度量学习的跨模态行人重识别训练框架，改善了光线较差情况下无法有效识别行人的问题。本文使用双流结构分别提取不同模态的行人图片，并对特征提取模块进行改造。在改造网络结构的基础上，本文提出跨模态双流困难三元组损失函数改进排序损失提升训练效果，以及使用焦点损失替代传统身份损失提升困难样本在训练中的学习权重。上述三种方法在 RegDB 数据集上进行实验并取得优异成绩。

(3) 搭建了一套监控视角下人物识别系统，验证了深度学习算法用于智能安防项目的可行性。系统搭建在多摄像头拍摄的室内环境中，使用 YOLOv3 目标识别算法和多重识别算法分别进行行人目标检测和人物识别。为了保证系统有效运行，本文不仅针对多重识别算法设计一套识别流程，还针对实时性需求设计多线程版本识别流程。系统经过实验收集并制作了单模态行人数据集为中层特征扩展方法提供实验数据，验证了人物识别系统的可行性。

关键词： 深度学习，行人重识别，表征学习，跨模态，困难三元组损失

DEEP LEARNING BASED PERSON RECOGNITION METHODS UNDER SURVEILLANCE CAMERAS

ABSTRACT

Person recognition under the perspective of surveillance is a key research direction in the field of intelligent security. Efficient and accurate person recognition algorithms can play an important role in scenarios such as public security criminal investigation that need to track specific targets. However, the pictures taken by the surveillance camera have complicated background information and noise interference, which makes traditional people recognition methods unable to effectively recognize them. Compared with traditional person recognition methods, deep learning methods have greatly improved accuracy and robustness in the field of person object detection and recognition. Therefore, person recognition methods based on deep learning have attracted a lot of attention.

This paper improves the single modality person re-identification and cross modality person re-identification from three aspects: convolutional network structure, ranking loss, and identity loss. Compared with the original method, the improved method has improved accuracy in the person re-identification dataset experiment. At the same time, based on the deep learning method, this paper designs a set of people recognition system from the perspective of monitoring, and provides a feasible solution for intelligent security projects. The main work of this article is as follows:

(1) A mid-level feature expansion method based on representation learning is proposed to alleviate the problem of overfitting the convolutional network model to the training data set in single modality person re-identification. In this paper, the convolutional network model is modified, and the middle-level features are extracted based on the backbone network for grouping and pooling operations, and then they are stitched with the high-level features to output. Experiments prove that the proposed method can effectively improve the generalization ability of the model and improve the accuracy of person recognition.

(2) A cross modality person re-identification training framework based on metric learning is proposed, which improves the problem that person cannot be effectively identified under poor light conditions. This paper uses a dual-flow structure to extract person pictures of different modalities, and transforms the feature extraction module. Based on the transformation of the network structure, this paper proposes a cross

modality dual-flow hard triplet loss function to improve the ranking loss to improve the training effect, and uses the focus loss instead of the traditional identity loss to improve the learning weight of the difficult samples in training. The above three methods are tested on the RegDB dataset and achieved excellent results.

(3) A set of person recognition system from the perspective of monitoring was set up, which verified the feasibility of deep learning algorithms for intelligent security projects. The system is built in an indoor environment with multiple cameras and uses YOLOv3 target recognition algorithm and multiple recognition algorithm to perform person target detection and person recognition, respectively. In order to ensure the effective operation of the system, this paper not only designs a set of recognition processes for multiple recognition algorithms, but also designs a multi-thread version recognition process for real-time requirements. The system collected and produced a single modality person dataset through experiments to provide experimental data for the middle-level feature expansion method, which verified the feasibility of the person recognition system.

KEY WORDS: deep learning, person re-identification, representation learning, cross-modality, hard triplet loss

目　录

插图清单

表格清单

第一章 绪 论

1.1 课题的研究背景及意义

随着科技发展以及经济水平提高，政府投入大量资金提升城市基础建设，期待构建智能安防及智慧城市。智能安防领域打破传统安防领域限制，通过与IT，电信等方面结合，使城市安防系统由点到面方向多方面的发展。在公安案件侦测场景下，人物特征识别方法与视频分析方法相结合能有效提升公安工作效率。人物特征识别目前常用技术为人脸识别，通过在大量实时视频中分析抽取人脸方式与公安库中人脸图像进行对比，公安能迅速准确得出目标人物并部署下一步行动，降低犯罪事件产生造成的巨大后果带来的影响。

由于智慧城市的需求以及数据时代来临，城市内的遍布大量功能各异的摄像头，如监控车流大小，拍摄违规开车行为等等。城市中数以万计的监控摄像头每天会产生PB级别的视频数据，大量的视频数据为我们生活带来便捷以及保障。然而在公安侦查场景下，大量视频数据带来繁重的工作量，警员通过人工方式处理由摄像头拍摄产生的大量实时视频数据已不现实，所以需要有一套成熟优秀的人物特征识别方法提高警员工作效率，提升监控摄像头在智能安防领域内的作用。

然而传统人物特征识别方法在由监控摄像头拍摄产生的视频下进行识别存在一定困难。首先，监控摄像头安装地点位于高处，拍摄角度一般带有俯视角度。其次，由于监控摄像头在高处进行拍摄，所以视频文件中存在复杂背景信息干扰并且人物在拍摄背景下所占比例小。人物特征识别方法需要准确的从复杂背景环境中分离人物，这一步是监控视角下人物识别的重要步骤，带有复杂环境信息的人物图片会严重影响识别效果。其次因为人物在拍摄背景下所占比例过小，分割小目标会存在困难。传统的分割方法包括基于阈值、基于边缘和基于区域的方法等[1,2]，这些传统分割方法对图片噪声干扰特别敏感，然而监控摄像头所拍摄图像和视频不可避免受各种不可控外部因素影响，无法产生高质量的图片。更重要的是人物识别是整个方法的核心，但当前常用的基于高阶特征的人脸识别方法[3]对人脸图片有严格的要求，需要正脸和全面的人脸特征信息，这在监控视角中是很难保证的。

在传统人脸识别方法存在困难的情况下，行人重识别方法被应用到改善监控视角下人物识别的问题上。行人重识别是在多摄像头网络下进行行人匹配的一种方法，在学术界里多强调其跨摄像头及无视野重叠的特点。对该方法的研

究始于多视频追踪[4]，并出现了基于图片的行人重识别[5]，基于视频的行人重识别[6]等分支。同时随着机器算力提升，以深层卷积神经网络为代表的深度学习方法广泛应用于计算机视觉任务中，包括图像分类，语义分割，目标检测等领域。基于深度学习方法强大的特征提取能力和分类能力，许多学者将其用于行人重识别方法中，并且在公开数据集中取得优秀效果。

本文基于深层卷积网络强大特征提取能力，在可将光模态行人重识别中基于表征学习存在模型过拟合问题提出中层特征改进方法，改进模型结构并在实验中取得较好效果。在可见光-热成像跨模态行人重识别中改进度量损失函数，训练跨模态卷积神经模型，并在公开数据集中达到领先水平。同时本文基于深度学习算法搭建一套监控视角下的人物识别系统，为智能安防中实际应用提供可靠方案。

1.2 国内外研究现状

监控视角下人物识别方法主要包含两部分方法，第一部分为目标检测方法，用于检测行人目标，第二部分为行人重识别方法，用于识别行人目标。基于深度学习的目标检测方法按照识别方式可分成单阶段检测算法和二阶段检测算法。单阶段检测算法通过在图中提取特征预测目标位置及分类，二阶段检测算法将首先生成候选区域提取区域特征，之后将所有区域特征放入一系列分类器中进行分类和精确定位。二者在识别时间及精度上各有优势，本文基于监控视角下人物识别系统的实时性需求，选择使用一阶段目标检测方法检测行人目标。行人重识别方法也分为两大领域，包括单模态行人重识别方法及跨模态行人重识别方法。单模态行人重识别方法解决在可见光模态下的行人识别问题，跨模态行人重识别方法解决在可见光-热成像或可见光-红外等多模态中行人识别问题。

1.2.1 目标检测方法研究现状

在深度学习尚未出现之前，目标检测领域使用人为构造几何特征进行特征提取并使用分类器对提取到的目标进行分类及位置修正，分类器包括 SVM 和 AdaBoost 等。常用人为构造几何特征包括 SIFT 特征[7]，HOG 特征[8]等。在 Hinton 提出 AlexNet 深度卷积网络[9]用于大规模图像分类后，深度学习方法广泛应用于目标检测领域，形成基于边框回归的单阶段目标检测方法及基于候选区域的二阶段目标检测方法。

一阶段目标检测方法直接将原始图像作为网络输入，将边框预测当成回归预测，通过卷积神经网络直接输出预测结果。基于边框回归的目标检测网络是一个端到端的网络框架，检测速度快，但是基于目标重叠较高时会出现误检情况。YOLO 目标检测算法[10]将输入图片划分为多个网格，每个网格负责预测在

该网格的目标个体，包括预测边界候选框以及类别预测概率。YOLO 目标检测算法将目标物体位置信息和类别信息同时输出，检测速度可以达到 45FPS，满足实时性需求。Redmon 等人随后提出 YOLOv2[11]及 YOLOv3[12]目标检测算法，在 YOLO 目标检测算法基础上使用更加深层卷积网络，引入批归一化层，锚点及多尺度预测，提升模型预测精度并仍然保持较快速度。一阶段目标检测方法由于其快速检测的特点，在实时追踪等具有实时性需求及检测准确度要求不高的场景下具有广泛的应用。但是在检测小物体的时候漏检率较高。

二阶段目标检测方法相比一阶段目标检测方法在检测速度上逊色不少，但是能保持较低漏检率。典型的基于候选区域的目标检测方法包括 Girshick 等人提出 R-CNN 目标检测算法[13]，以及 Faster R-CNN 目标检测算法[14]等。R-CNN 算法通过选择性搜索方法产生候选区域，并将所有候选区域输入至卷积网络中提取特征，最后通过支持向量机分类器进行分类及线性回归修正边框。R-CNN 缺陷在于整个框架不是一个端到端网络，提取特征和分类是分开进行，效率较低。Faster R-CNN 使用 RPN 网络替换 Selective Search 方法生成候选区域，实现一个端到端网络框架，提升训练效率，但是在检测速度上仍然无法达到实时性需求。二阶段目标检测方法由于采用多步骤识别的特点，在检测准确度上具有明显优势，适用于需要精度要求高的复杂场景。

1.2.2 可见光模态行人重识别研究现状

行人重识别是在多摄像头网络下进行行人匹配的一种方法，在学术界里多强调其跨摄像头及无视野重叠的特点。可见光模态行人重识别所识别的图片都是由可见光摄像头进行拍摄，整个数据集中仅存在可见光模态的图片。传统可见光模态行人重识别方法有基于低层视觉特征[15]，基于中层语义特征[16]以及基于高层抽象特征[17]等。这些传统行人重识别方法提取特征有限，抗干扰能力差，很难在大量视频数据中保持准确率。

近年来，深度学习由于其强大的自动特征提取能力和分类能力，被广泛运用在计算机视觉领域内。目前基于深度学习的行人重识别方法在某些公开的数据集上已经有了非常高的准确率，如 Li 等人[18]提出均衡注意力卷积神经网络（Harmonious Attention CNN），Sun 等人[19]提出基于区域卷积结构的卷积神经网络等。基于深度学习的行人重识别方法主要包括两个方面，分别是基于表征学习的行人重识别方法和基于度量学习的行人重识别方法。表征学习方法将行人重识别任务当成分类问题看待，利用行人类标或者行人属性等作为标签训练模型，让网络学习这两张图片是否属于同一个人。表征学习由于训练数据集存在不够全面的特点，模型训练过程中会出现对训练集过拟合问题导致训练出来的模型泛化能力不强，所以需要对网络结构以及训练方法上改进缺陷。度量学习方法主要是通过网络学习提取两张图片的相似度，同一个类标的行人的不同图片的相似度要大于不同类标的行人图片。度量学习方式由于使用相似度计算

损失，当损失函数参数定制不恰当时，训练过程中会出现损失无法收敛的情况，造成训练资源浪费。

1.2.3 跨模态行人重识别研究现状

基于可见光模态下行人重识别方法在数据集上已经拥有很高准确率，但是在实际场景下表现效果不好问题，研究人员认为在实际场景下存在多种光照条件，不同光照条件拍摄图片的模态间差异会造成模型准确率下降过大，所以提出跨模态行人重识别问题。跨模态行人重识别研究包括可见光-深度跨模态行人重识别（RGB-D Re-ID）[20]，可见光-热成像跨模态行人重识别（VT-REID）[21]以及可见光-红外跨模态行人重识别（RGB-IR Re-ID）[22]。RGB-D Re-ID 用于匹配 RGB 图像和深度图像间的人物信息，深度信息提供不变的身体形状和轮廓信息并且不受光线和颜色变化影响。VT-REID 和 RGB-IR Re-ID 用于匹配 RGB 图像和红外图像间的人物信息，不同的是 VT-REID 中的红外图像采用红外热成像技术获取热成像图像信息，RGB-IR Re-ID 的红外图像采用红外摄像机主动发射红外光并通过自带收集器收集返回红外光并获得相关红外图像。跨模态行人重识别研究较少。Wu 等人[22]首次提出 Zero Padding 方法，并发布 RGB-IR 数据集 SYSU-MM01。Nguyen 等人[23]收集并提出 VT 数据集 RegDB。跨模态行人重识别方法由于尚未成熟，在公开的跨模态行人数据集中准确度较低，所以还需要研究人员继续深入研究，争取提出更加优秀的算法提高准确率。

1.3 本文主要研究工作

首先，针对可见光模态中行人重识别算法在公开数据集中达到较高水平，但是在实际情况下无法达到较好效果的情况，提出中层特征扩展方法，增加卷积网络模型特征表征能力。其次基于跨模态行人重识别中多模态的特点，使用改进特征提取模块的双流结构分别提取不同模态的行人特征。由于度量学习在可见光模态行人重识别模型训练的有效性，本文通过改造困难三元组损失，使其困难采样范围扩大至多个模态，以此提升跨模态行人重识别中模型识别准确率。在改造损失函数基础上，基于可见光−热成像跨模态行人重识别中红外模态有效信息少问题，本文使用焦点损失替换交叉熵损失进行模型训练，减轻可见光−热成像跨模态行人重识别中由于红外模态图片难以识别造成的样本失衡问题。最后本文基于室内环境搭建一套监控视角下人物识别系统，利用该系统拍摄大量有效图片制作成数据集用于验证可见光模态下行人重识别卷积网络模型的在实际场景下的识别效果。本文全文工作可以分为以下几个部分：

（1）设计了一种神经网络中层特征提取方法，用于提升模型泛化程度；由于可见光模态行人重识别数据集中场景单一，光线差异不明显，训练过程中容易造成模型过拟合情况。针对这种情况，本文基于中层表征特征蕴含丰富信

息，对模型的第 3、第 4 层进行改造，提取中层特征经过拼接池化等操作，提升模型泛化能力，并在验证数据集中达到较好效果。

（2）提出了多种改进方法的跨模态行人重识别模型训练框架，用于提升模型在跨模态行人重识别中准确率；跨模态行人重识别数据集相对于可见光行人重识别数据集最显著差异在于它的多模态。跨模态行人重识别模型往往会出现多个不同模态情况，如可见光-深度跨模态行人重识别，可见光-热成像跨模态行人重识别等，所以本文使用双流结构分别提取不同模态行人图片的特征。为了让模型更好学习相同行人图片在跨模态中相似表征特征，本文对可见光模态的传统困难三元组损失进行改造，将困难样本选择扩展至多模态中。其次为了进一步提升跨模态卷积网络模型的识别能力，本文使用焦点损失函数使模型在训练阶段注重难样本图片的学习。通过公开数据集 RegDB 验证，以上方法均具有提升模型识别准确率的能力。

（3）建立了室内可见光模态监控视角下人物识别系统，用于制作可见光模态的实验数据集和提供针对智能安防人物识别的可行方案；由于可见光模态行人重识别数据集中存在图像中光线及视角大量一致情况，导致训练模型运用至实际场景中会出现不同程度的准确度下降问题。针对这种情况，本文为了验证扩展中层特征改造模型在实际场景下有效性，在室内空间下采用多个摄像头不间断拍摄收集 11 天视频，并使用 YOLOv3 目标检测算法提取有效目标 7610 个，并进行标注，制作成验证集用于验证模型在实际场景下准确性。同时设计了单线程以及多线程识别流程，为智能安防项目提供一套可行方案。

1.4 本文组织架构

本文重点为可见光行人重识别研究，跨模态行人重识别研究以及监控视角下人物识别系统，共分成六个章节进行描述，整体安排如图 1-1 所示：

第一章，绪论；主要介绍了行人重识别领域研究的背景意义以及现状等，分析了传统人物识别方法在监控视角下的缺陷，引出使用行人重识别方法克服监控视角下人物识别难点，并简单介绍行人重识别领域内传统人工特征方法以及深度卷积网络方法的发展。

第二章，相关技术理论；主要对本文使用的深度学习技术进行简单介绍以及相关实用方法描述。

第三章，监控视角下基于可见光模态行人重识别方法研究：本章节设计一种深度卷积网络中层特征扩展方法，通过对深度卷积网络模型进行改造，使模型在公开数据集上进行训练并获得较强的泛化能力。最后通过由自建监控视角下人物识别系统拍摄的图片作为数据集进行实验，验证网络改造方法的有效性。

第四章，基于可见光-热成像模态行人重识别方法研究；本章节提出基于多种改进方法组成的训练框架，用来进行可见光-热成像模态人物识别。最后通过对比实验证明方法有效性。

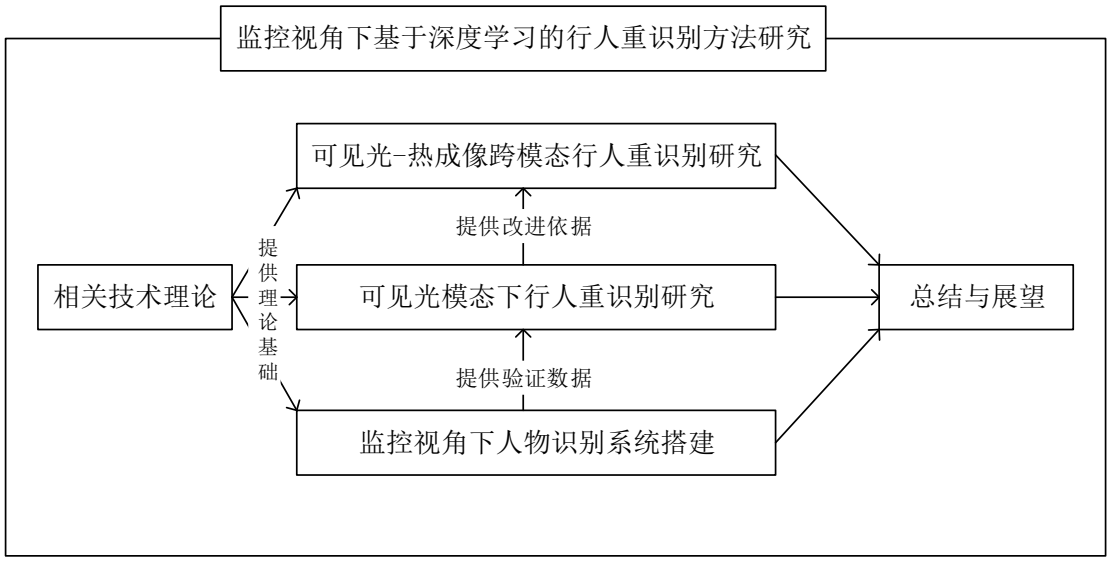


图 1-1 章节安排图

Figure 1-1. Chapter arrangement figure

第五章，监控视角下人物识别系统搭建；本章节提出一种基于深度学习方法的人物识别系统，通过在实验环境中搭建多摄像头网络进行拍摄，并经过目标检测算法剪切出行人图片。获得行人图片后使用行人重识别算法将图片与标准库中进行匹配，最终达成识别目的。

第六章，总结与展望；本章节对本文内容进行总结，分析本文提出中层特征扩展方法以及跨模态行人重识别框架的优点以及不足，同时对存在的问题和以后的研究工作进行思考和展望。

1.5 本章小节

本章节首先介绍了课题研究背景及意义，同时对国内外行人重识别相关研究现状进行分析及讨论。针对现存行人重识别算法在实用场景下准确率较低问题，提出使用中层特征扩展方法对卷积网络模型进行改造增加模型泛化能力，并在由自建的监控视角下人物识别系统拍摄整理而成的数据集中验证有效性。其次针对行人重识别领域中新兴的可见光-热成像领域，提出基于改进困难三元组损失框架训练模型，最终在对比实验和消融实验中证明框架有效性。最后介绍了本文各个章节安排。

第二章 深度学习技术理论

深度学习学科近年来发展迅速，引起国内外广泛的关注，并在学术界和工业界取得了重要突破和大量应用。深度学习首次浪潮出现在20世纪50年代末，当时一种简单的感知器神经网络模型出现引起了大量科学家对神经网络兴趣，但是由于感知器不能解决简单非线性问题，所以很快进入第一个寒冬期。在1986 年辛顿提出 BP 算法让人工神经网络解决非线性问题，使其神经网络重新回到视野中，但是由于出现梯度消失及硬件算力不足问题，第二个寒冬期也来临。21 世纪以来，随着硬件算力提升以及无监督和有监督学习算法结合应用，卷积网络再次吸引大量科学家研究。

深度学习是卷积网络基于硬件算力提升和训练算法不断优化而出现的新兴领域。深度学习通过无监督特征学习和分层特征提取来替代手工获取特征，并在图像分类，目标检测，自然语言处理等领域内获得成功。本章节主要介绍本文使用深度学习的一些基础知识，训练方法以及相关应用等等，希望通过本章的介绍可以理解本文的主要内容。

2.1 卷积神经网络

卷积神经网络研究最早起源于日本学者 Fukushima 提出的 Neocognitron 模型[24]，这个模型是一个具有级联结构的卷积神经网络，网络按照 Simple-layer 层和 Complex-layer 层进行重复堆叠，最终输出的是 0-9 的分类结果。Neocognitron 通过 Simple-layer 层提取局部特征并输入至 Complex-layer 层，完成底层局部特征到高层抽象特征整合，被认为是卷积神经网络的开创性研究。在之后的研究中 Yann LeCun 在 1998 年提出 LeNet5 卷积神经网络[25]被认为是现代卷积神经网路雏形，LeNet5 网络模型相较于 Neocognitron 网络模型关键不同在于 LeNet5 使用 BP 反向传播算法进行有监督学习，并使用池化层和激活函数等方法提升模型准确率。但是由于硬件计算能力限制，在一般的实际任务中表现不如支持向量机，Boosting 等方法好。直到 2012 年，Hinton 提出的 AlexNet 网络[9]在 ImageNet 图像识别大赛中将识别误差率降低至 15%时，卷积网络才重新被广大研究人员认可，并广泛运用至学术界和工业界中。

卷积神经网络广泛应用主要有以下方面的原因：（1）局部感知，早期多层感知器中，隐藏层的所有节点都连接到输入图像或者特征图的每个像素点上，而在卷积神经网络中，每个隐藏层节点只连接到输入图像或特征图中一部分足够小的局部像素点上，从而大大减少需要训练的权值参数；（2）权值共享，不

同的输入图片或者同一张图片的不同局部相聚点共用一个卷积核，减少卷积核数量。原因在于不同输入图片当中可能会出现相同的特征，共享卷积核能够进一步减少权值参数。（3）池化，将输入图片或者特征图经过卷积核卷积后，都通过一个池化过程，来减小图像的规模。池化好处在于缩小图片和特征图的规模，提升计算速度，同时减少下一层卷积核的参数，防止卷积神经网络出现过拟合现象。基于上述的优点，卷积神经网络才能在近年来大放异彩，成为深度学习等新兴技术的底层基础。

2.1.1 卷积神经网络基本结构

卷积神经网络经过大量科学家研究及优化，普遍包含 3 个基本结构，分别是卷积层，池化层以及全连接层。

卷积层是由多个特征图组成，其中特征图中的每个元素值为卷积核的输入。当以图片作为卷积网络的输入时，图片的通道值则为卷积核输入，图片中的通道值经过卷积核计算后，输出为特征图中的元素值。特征图中的元素值可以作为下一个卷积核的输入，最终输出结果为下一个卷积层的元素值。

卷积核是卷积神经网络用来提取特征的工具，每一层卷积层可以由多个卷积核进行特征提取。卷积核层次越高，提取的特征更加具体。卷积核是一个 n × n × d的权重矩阵 W，矩阵中的元素为卷积核的权重 w，其中 n 为卷积核的大小，d 的大小依据输入数据的维度判断，若输入数据是黑白照片则 d 值为 1，若输入数据是彩色图片则 d 值为 3。另外每个卷积核中还有偏置 b，用于进行卷积运算。卷积运算过程如图 2-1 所示，其计算公式如(2-1)所示，矩阵形式如(2-2)所示。

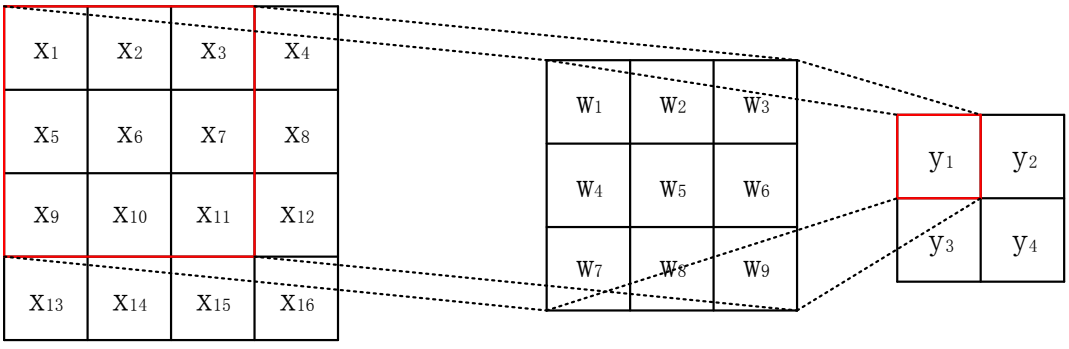


图 2-1 卷积计算示例

Figure 2-1. Convolution calculation example

$$y1 = (\sum_{i=1}^{9} x_i w_i) + b \tag{2-1}$$

$$Y = WX + b \tag{2-2}$$

在卷积神经网络中卷积核大小远小于输入数据或特征图大小，卷积核需要对输入数据或特征图依次滑动提取局部特征，并最后组成一个输出特征图。卷积核滑动遍历过程还需要注意两个参数，（1）卷积核滑动步长，滑动步长规定卷积核相邻两次划过特征图的距离，步长为 n 时，卷积核在一次扫描后跳过 n-1 个像素点。（2）填充方式，由于卷积核滑动过程中不可能出现每次都能刚好划完情况，需要进行填充操作，如有效填充，半填充等。这两个参数根据实际需求具体选择。如下所示，当输入卷积层大小为5×5，卷积核大小为3×3时，卷积步长为1×1及填充为 0 时，输出卷积层大小为3×3。当输入卷积层大小为5×5，卷积核大小为3×3时，卷积步长为2×2及填充为 1 时，输出卷积层大小为3×3。两种参数卷积示意图如图 2-2 所示。

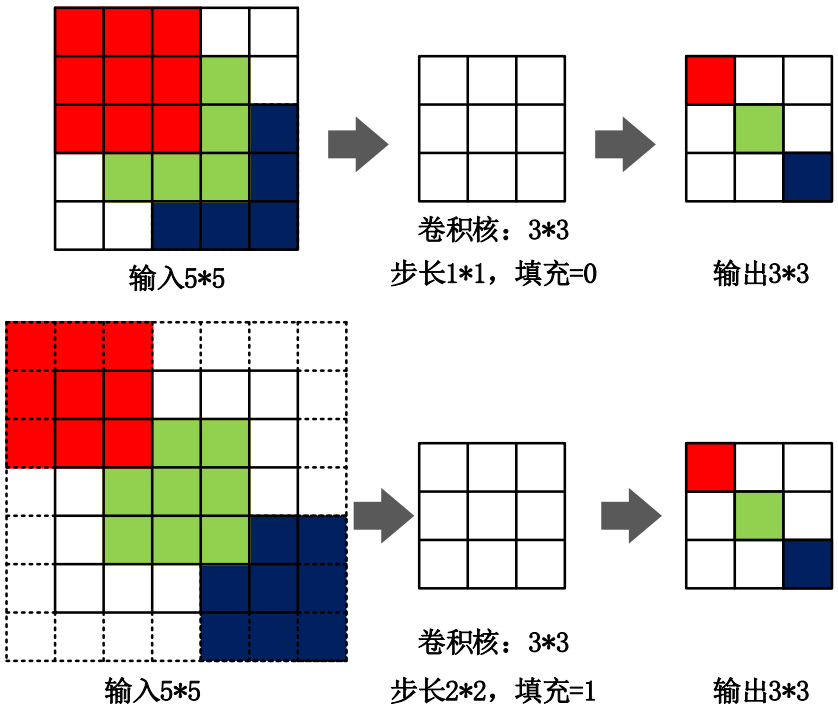


图 2-2 带参数卷积计算示例

Figure 2-2. Convolution calculation example with parameters

卷积神经网络里上一层网络的输出就是下一层网络的输入，上一层特征图元素与下一层特征图元素是以卷积核进行连接，每个卷积核需要滑动整个特征图提取全部的局部特征，并组合成下一层特征图。这种做法称为权值共享，降低了算法复杂度，提升训练速度。

池化层又称为下采样层，池化层在卷积神经网络结构中通常位于卷积层后，池化层由特征图堆叠而成，池化层中的元素连接卷积层部分的元素，每个元素把连接的卷积层元素作为输入，经过特定操作输出处理后的元素值。因为经过卷积核获得的卷积层含有大量数据以及无效信息，所以需要进行操作剔除无效信息并降低数据量。池化层主要作用是通过减少特征图的大小和维度以减

少计算复杂度。池化层主要分为以下 3 种：（1）最大池化；（2）平均池化；（3）随机池化。

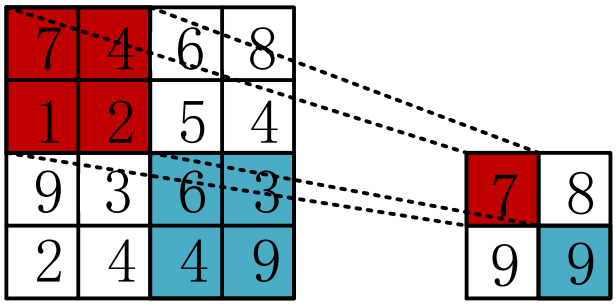


图 2-3 最大池化示例图

Figure 2-3. Maximum pooling example graph

全连接层通常位于卷积神经网络中所有卷积层之后。全连接层中的元素需要与输入特征图中的所有元素相连接，所以全连接层会占用大量空间。全连接层主要作用是对经过多个卷积层和池化层后的特征图提取高层特征，高层特征具有明显区分性，可以实现分类目标。通常全连接层后会使用 softmax 函数输出图片分类结果，训练过程中还可以使用优化操作提升模型训练效果。

2.1.2 卷积网络优化结构

对于图片来说，前一节提到使用卷积层进行处理，对输入图片或者特征图中的元素使用卷积核提取局部特征，这些是线性操作，但是对于样本及目标来说，会出现非线性可分的情况。为了解决非线性可分的困难，常用方法为加入非线性因素，即激活函数层。

激活函数层提高卷积神经网络对数据的表达能力，解决线性模型不能解决的非线性问题。常用的激活函数有 Tanh 激活函数，Sigmoid 函数，ReLU 函数 3 种。三种函数公式如下：

$$\mathrm{Tanh}(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}} \tag{2-3}$$

$$\mathrm{Simoid}(x) = \frac{1}{1 + e^{-x}} \tag{2-4}$$

$$\mathrm{Relu}(x) = \max(0, x) \tag{2-5}$$

如图 2-4 和图 2-5 所示，Tanh 激活函数，Sigmoid 函数在接近饱和时梯度值非常小，容易导致出现梯度消失问题。因为 BP 算法进行反向传播的时候，梯度是以乘法方式由高层传递至底层，当层数较多时，若出现梯度值非常小的情况，网络权值则不能有效更新，造成训练低效。Relu 函数在$x > 0$时梯度为恒定值，无梯度消失问题，模型收敛速度快，其次 Relu 函数还增加了网络稀疏性，

当x＜0时，输出为 0，当训练结束后，由于特征图稀疏性大，所以提取出来的特征就具有代表性，卷积神经网络模型泛化能力则越强。

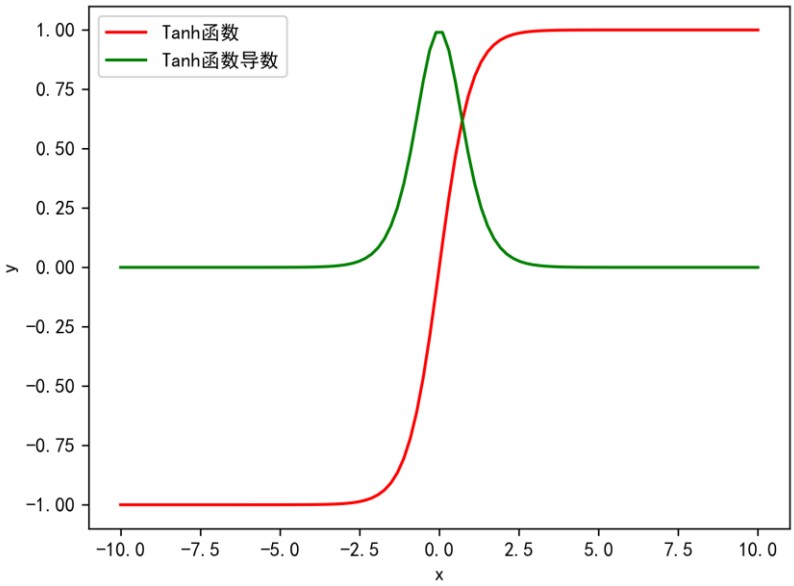


图 2-4 Tanh 函数及导数图

Figure 2-4. Tanh function and derivative graph

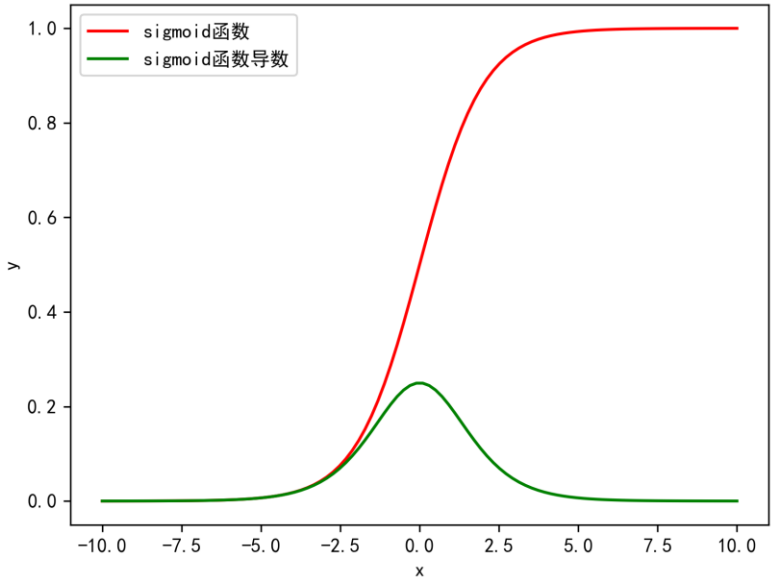


图 2-5 Sigmoid 函数及导数图

Figure 2-5. Sigmoid function and derivative graph

批量归一化层是 2015 年提出的一种优化训练的方法，在没提出批量归一化层时，通常使用需要人为选择参数的梯度下降法。梯度下降法训练效果非常依赖于参数选择，导致研究人员在调参上浪费大量时间。批量归一化层则解决了对于参数选择的依赖，提升训练速度，加快模型收敛。归一化是一种数据预处理方法，在卷积神经网络模型训练开始前，对数据做归一化处理可以限定输入数据分布，增加网络的泛化能力。由于输入数据经过网络训练时，网络每层的

参数的分布也是发生变化的，所以为了防止网络内部中因为数据分布问题导致训练低效收敛过慢的问题，批量归一化层被提出用于解决网络内部数据分布变化问题。批量归一化公式如下：

$$\mu_\beta = \frac{1}{m}\sum_{i=1}^{m} x_i \tag{2-6}$$

$$\sigma_\beta^2 = \frac{1}{m}\sum_{i=1}^{m}(x_i - \mu_\beta)^2 \tag{2-7}$$

$$\hat{x}_i = \frac{x_i - \mu_\beta}{\sqrt{\sigma_\beta^2 + \varepsilon}} \tag{2-8}$$

$$y_i = \gamma\hat{x}_i + \beta \tag{2-9}$$

2.1.3 损失函数

损失函数在卷积神经网络训练中具有巨大作用，它有助于优化卷积神经网络参数。卷积神经网络训练的目标是最小化模型预测值与数据实际值的误差，所得到的误差用来计算损失并对模型进行反向传播训练。卷积神经网络在训练时如果预测值与实际值的差距越大，那么在反向传播训练时，模型参数调整幅度则需要更大，从而使训练更快收敛。损失函数有以下几种常见损失函数，分别是平方误差损失函数，交叉熵损失函数以及焦点损失函数。

平方误差损失函数是简单的回归损失函数，它的平方误差损失是实际值和预测值之间差值的平方。二次函数仅具有全局最小值，所以可以保证模型训练收敛至最小值。但是平方误差损失对异常值敏感，如果数据中存在许多异常值时，则不应该使用该损失函数。平方误差损失函数公式如下：

$$SL = (y - f(x))^2 \tag{2-10}$$

交叉熵损失函数是分类问题中最常用的损失函数，交叉熵能表示同一个随机变量中两个概率分布的差异值，在深度学习中被认为表示模型预测概率分布和数据真实概率分布的差异值，交叉熵值越小，模型预测效果则越好。交叉熵损失函数经常和 softmax 函数配合使用，先使用 softmax 函数将模型对多分类的预测值之和整合为 1，之后通过交叉熵计算损失。交叉熵损失函数公式如下：

$$H(p,q) = -\sum_{i=1}^{n} p(x_i)\log(q(x_i)) \tag{2-11}$$

焦点损失函数[26]是交叉熵损失的改良版本，它主要解决一阶段目标检测中正负样本比例失衡问题。它主要的创新在于在计算过程中降低简单负样本在训练中占有权重，提升对困难样本的学习。焦点损失函数思想和困难样本挖掘相似，但是它只是降低简单样本权重而不是像困难样本挖掘一样剔除容易样本学

习权重。焦点损失函数在一阶段目标检测算法中效果明显，但是利用到跨模态行人重识别领域内同样效果明显。焦点损失函数公式如下：

$$FL(p_t) = -(1-p_t)^{\gamma}\log(p_t) \tag{2-12}$$

2.2 目标检测算法

目标检测任务与图像分类任务存在本质上的不同，图像分类着重于图片的整体，输出结果是整个图片的描述，而目标检测则关注图片局部特定的物体目标，输出的结果是特定目标在整体图片中的位置信息以及类别信息。目标检测需要从背景中分析出特定目标，并确定这一目标的描述，所以模型输出结果是一个综合数据块，数据块中每一项中分别有目标的类别信息和矩形检测框的位置信息。目标检测算法目前广泛应用于各类实用型产品中，效率和准确率也日益提升，目标检测算法主要是按照处理方式进行分类。

2.2.1 目标检测算法处理方式

目标检测算法按照处理方式可以分成二阶段目标检测算法和一阶段目标检测算法，经典的目标检测算法按照发布时间排序如图 2-6 所示。简要来说，二阶段目标检测算法首先由算法生成大量候选框，再通过卷积神经网络进行分类。一阶段目标检测算法则跳过候选框阶段，直接将确定目标边框问题转换成回归问题进行处理。

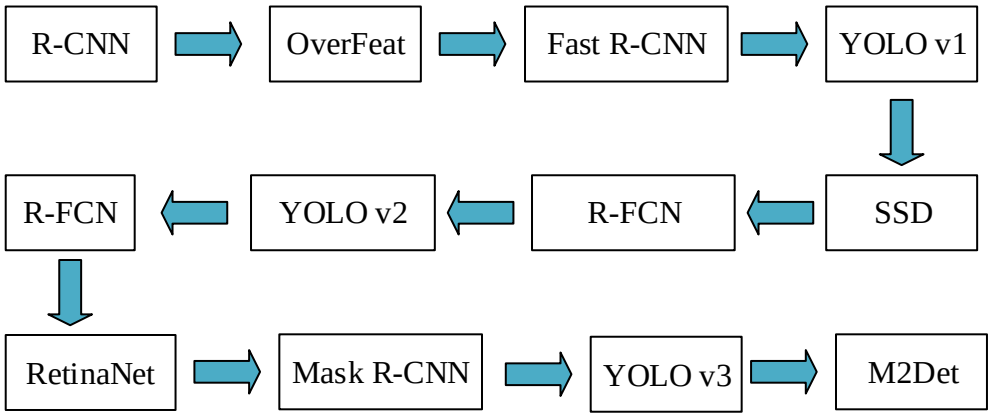


图 2-6 经典目标检测算法示意图

Figure 2-6. Schematic diagram of classic detection algorithm

二阶段目标检测算法中经典算法是 R-CNN 系列目标检测算法。2014 年，Girshick 等人提出 R-CNN 目标检测算法[13]，该算法利用选择性搜索算法评估出较可能的区域作为卷积神经网络输入，让卷积网络学习候选框和标定框组成的正负样本特征，最后由分类器分类后对候选框及边框完成回归操作完成目标定位。Faster R-CNN[14]为了减少生成候选框造成的速度过慢问题，提出了区域建议网络，使整个卷积神经网络流程都能共享卷积神经网络提取的特征信息。目

前 R-CNN 系列已经升级至 Mask R-CNN 目标检测算法[27]，在准确度上达到优异效果。

一阶段目标检测算法在检测精度上无法达到像二阶段目标检测算法的准确性，但是在某些实时性场景下，二阶段目标检测在速度上无法满足要求。一阶段目标检测算法采用基于回归的方法实现单次训练中共享特征，保证准确率前提下提升检测速度。常见的一阶段目标检测算法包括 YOLO 系列算法以及 SSD 算法[28]。SSD 算法将 YOLO 回归思想和 Faster R-CNN 的锚点机制结合，在保持快速性的同时准确度还有不小提升。

2.2.2 YOLO 系列目标检测算法

YOLO 算法[10]继承 OverFeat 算法[29]的基于回归的一阶段目标检测算法，通过对图像的全局信息进行预测，整体结构简单。训练阶段首先将图片大小重新调整至448×448固定大小，并在此基础上将图片划分为7×7的网格区域，经过卷积网络训练直接预测每个网格内的矩形边框信息和类别的置信度。这种方式速度极快，能达到每秒 45 帧的处理速度，但是存在矩形边框定位信息不够准确以及在目标检测领域内的指标不如二阶段目标检测方法，且对于小物体检测效果较差。

Joseph Redmon 等人于 2017 年提出 YOLOv2[11]作为 YOLO 的改进版本，重点解决了矩形边框定位精度不足和召回率低的问题。YOLOv2 选择 Darknet-19 卷积网络作为特征提取网络，并在图像训练初始阶段加入批量归一化的预处理。在此基础上，YOLOv2 还使用224×224及448×448两种固定图片尺寸的方式微调 ImageNet 预训练模型，提升模型检测能力。YOLOv2 引入锚点机制，采用 K-Means 聚类方法训练出锚点模板，并在卷积层使用锚点框操作和采用多限制条件的定位方法，极大提升算法召回率，同时有利于小尺寸目标检测。

YOLOv3[12]则是最新的 YOLO 目标检测系列算法，在 YOLOv2 的基础上，YOLOv3 采用了 Darknet-53 的深层卷积网络结构，参考了残差网络的思想，在一些卷积层之间设置了一些快捷链接通道。其次，YOLOv3 采用了多尺度特征进行对象检测并增加了多种尺度的先验框，提升小物体检测效果和 mAP 指标。在一些细节上，YOLOv3 放弃使用 softmax 函数，而是使用独立逻辑分类器，使其适用于多标签分类。本文中提及的人物识别系统使用的目标检测算法的也是 YOLOv3 目标检测算法。

2.3 行人重识别算法

行人重识别算法是近年来学术界热门的研究领域，大量研究人员不断提出新的方法并在公有数据集上达到较高准确率。行人重识别算法是在多个视角不

重复的摄像头网络下进行行人匹配的过程。行人匹配就是在多个摄像头在不同时间拍摄的多个行人目标中寻找是否存在与锚点行人身份相同的行人目标。目前行人重识别领域主要研究是基于可见光模态下的行人目标识别，但是最新研究中已经出现不少跨模态行人重识别问题，如可见光-深度跨模态行人重识别问题，可见光-红外热成像跨模态行人重识别问题以及可见光-主动红外跨模态行人重识别问题。这些领域覆盖了光照充足与不充足的情况下的行人匹配，有利于智能安防产业在所有时间段内发挥重要作用。所以行人重识别问题吸引了大量学者研究，目前比较常见的方法包括基于表征学习以及基于度量学习等方法。

2.3.1 表征学习

表征学习是深度学习中常用的一种方法，其关键在于如何找到具有辨别性特征。常用的特征包括底层手工特征，中层语义特征，高级抽象特征。底层手工特征大部分都是对图片切分成多个部分，对每一个部分提取底层特征组合表示成具有辨别性的特征表示。最常见的方法包括基于颜色特征的 HSV 直方图表示，基于形状特征的方向梯度直方图[30]等等，这些方法属于行人重识别中传统方法，在早期研究中使用广泛，但是准确率不高。中层语义特征是通过语义信息判断不同图片中的行人目标是否属于同一个行人。Layne 等人[16]在 Market-1501 和 DukeMTMC-ReID 公开数据集上手工标注多达 20 多种属性的语义信息来描述行人目标，包括性别，上装颜色，头发长短，是否佩戴帽子等信息，经过多种对比实验证明中层语义特征的有效性，在行人重识别的各类指标中有大量提升。高级视觉特征在特征处理上比底层视觉特征繁琐，在底层特征分割图片基础上，对一些重要纹理信息进行 Fisher 向量编码[31]等，或者对局部块使用高斯分布建模[32]等方式获取高级抽象特征。

随着深度学习发展，卷积神经网络也被用于提取不同图像中行人目标的特征。由于卷积神经网络具有依据任务自动将低级视觉特征组合成高级抽象特征的特点，许多研究人员将深度学习图像分类的方法应用至行人重识别问题中，利用行人的类别信息作为训练标签训练卷积网络模型。为了增加卷积神经网络对于行人目标的表征能力，许多研究人员对网络提出各种改进，Li 等人提出为卷积网络构建基于区域和基于像素两种注意力结构，更深层次提取图片中辨别性信息[18]，Sun 等人提出在没有标注情况下将行人外观特征分解为多个具有辨别性语义层次，让卷积网络从多种语义中提取特征并组合成更加具有辨别性特征[19]。表征学习方法是行人重识别中较为常用的方法，但是容易存在数据过拟合问题，往往只对训练数据集的场景下具有效果，在其他场景下效果不明显。

2.3.2 度量学习

相比表征学习中学习出具有辨别性特征的方法来说，度量学习方法将行人图片转换成度量空间中的空间点，选取任意一张行人图片作为锚点图片，让相

同行人类别的不同行人图片和锚点图片在度量空间中的距离尽可能小，让不同行人类别的所有图片和锚点图片的在度量空间中距离要尽可能远。基于度量空间距离的思想，Schroff 等人提出三元组损失[33]，Varior 等人提出对比损失[34]，Chen 等人提出四元组损失[35]等等。

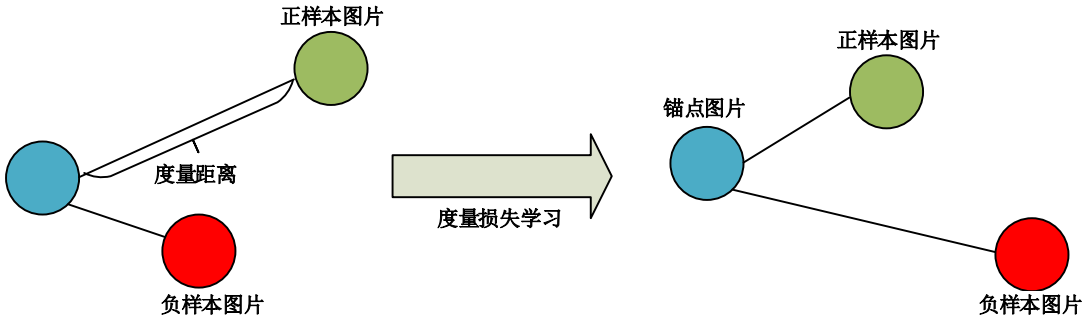


图 2-7 三元组损失示意图

Figure 2-7. Triplet loss diagram

三元组损失是度量学习方法中广泛运用的一种方法，从一次训练的行人图片中，选取 3 张图片作为一个三元组，其中一张图片为锚点图片，另两张图片分别为锚点图片的相同行人类标的图片以及不同行人类标的图片，三元组损失计算公式如下：

$$L_T = max\left(d_{a,p} - d_{a,n} + \alpha, 0\right) \tag{2-13}$$

基于三元组损失的改进有很多，其中有两种改进方法被运用的较多。一种改进方法是将三元组损失中考虑的正负样本对之间的相对误差变为绝对误差，进一步将正负样本对的距离拉开[36]。另一种改进方法是基于困难样本挖掘，对于一批量的行人图片，选择一张图片作为锚点图片，在与锚点图片行人类标相同的图片中选择与锚点距离最大的行人图片，在与锚点图片行人类标不同的图片中选择与锚点距离最小的行人图片，组成困难三元组损失[37]。困难三元组损失训练的模型在公开数据集中效果相比传统三元组损失存在较大提升。改进三元组损失和困难三元组损失公式如下：

$$L_{IT} = d_{a,p} + max\left(d_{a,p} - d_{a,n} + \alpha, 0\right) \tag{2-14}$$

$$L_{IT} = \frac{1}{P * K} \sum_{A \in batch} \max(\max_{p \in A} d_{a,p} - \min_{n \in B} d_{a,n} + \alpha, 0) \tag{2-15}$$

2.4 本章小结

本章主要介绍了本文使用的相关技术的基本内容。卷积网络是深度学习的基础，没有强大的硬件算力以及优秀的训练算法支撑，深度学习无法成为近年来火热的领域。卷积网络强大的特征提取能力是其能被运用在各个计算机视觉

领域的原因。目标检测算法通过回归预测或候选框分类回归方式进行目标检测并各有优劣。行人重识别算法通过表征学习和度量学习方式等逐渐完善并在公开数据集上达到优异效果。

第三章 监控视角下基于可见光模态行人重识别方法研究

监控视角下的人物识别具有无法有效提取行人目标和难以准确识别行人的问题，同时使用单一数据集训练的卷积网络模型会存在对训练数据集过拟合问题。面向当前监控视角下人物识别的困难，本文提出一种基于深度学习方法的行人重识别方案。该方法首先针对深度学习模型泛化能力不强的问题，通过对模型的输出进行改造，单独对网络模型的中层输出使用不同池化策略，有效地提升网络模型的跨域匹配能力；另外提出了一种动态行人库方法，在视频信息识别的过程中不断将达到给定阈值的图片更新到标准库中，提升标准库下时间信息的获取，进一步提升识别算法精度。方案的有效性在自建数据集和 Duke-MTMC 数据集上得到了验证。

3.1 人物识别方法及步骤

所提出方案的整体设计如图 3-1 所示。首先，在人物定位阶段，针对原始图像中存在干扰信息较多的问题，使用 You Only Look Once（YOLOv3）目标检测算法对原始图像中的人物进行定位，从复杂的背景环境中提取出仅包含少量背景信息的人物图片，实现复杂背景信息和固有噪音的剔除。其次，在人物识别阶段，将上阶段得到的处理后的人物图片输入到改造后的卷积网络模型中提取出图片特征，并与标准库中的特征进行对比，得出识别结果。最后使用动态行人库的方法，将度量小于规定阈值的图片作为标准库的图片加入到库中，提升识别能力。

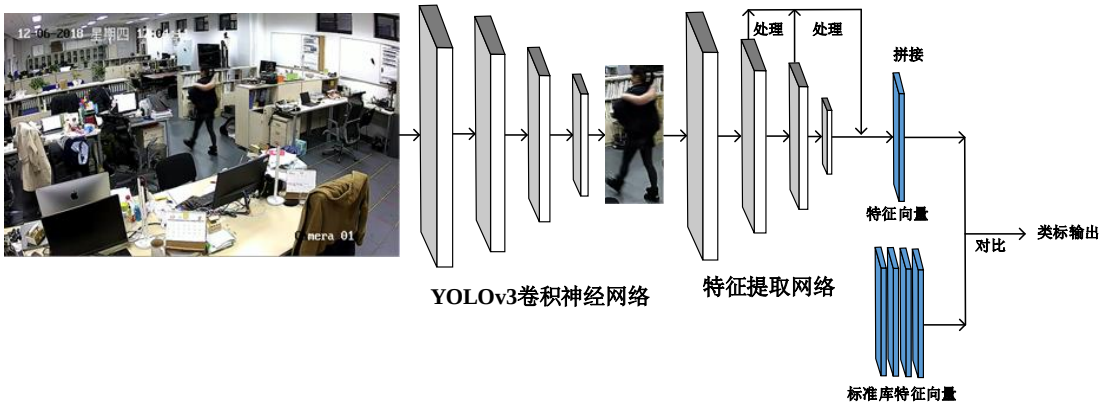


图 3-1 人物识别流程图

Figure 3-1. Person recognition process diagram

3.1.1 人物对象的定位

监控视角下获取的视频信息包含了复杂的背景环境，光照变化等干扰，特征提取器对无效信息干扰非常敏感，会极大影响提取器的性能，因此如何从不确定的环境中精确地分割出人物是极其重要的一步。

现有目标检测算法已经较为成熟，考虑到处理视频数据时的快速性要求，本文采用基于 YOLOv3 的目标检测算法进行人物检测。YOLOv3 是由 Redmon 等人[12]提出来的一套改进的目标检测算法，具有快速，背景误检率低，通用性强等优点，适合方案目标。

3.1.2 人物特征提取及对比

本文将行人重识别看成分类问题，将行人图片放入卷积神经网络模型中进行特征提取并与标准库中图片的特征进行距离对比，最后将测试图片的类标与标准库中距离最相近的特征类标归为一类，达成识别的目的。

本文使用 Resnet50 作为骨干网络，通过对网络结构的修改，使修改后的网络模型具有较强的跨域识别能力，提升网络模型对于监控视角下行人图片的特征提取性能。在网络改造的基础上，采用迁移学习的思想，将修改过的网络模型放入 Market1501 行人重识别数据集中进行训练，通过选取 adam 优化算法及交叉熵损失函数和 Tri-Hard 损失函数来达到更好的训练效果。

进行特征提取的步骤如下：（1）将待测图片$\boldsymbol{M}$作为 YOLOv3 卷积神经网络模型的输入，网络输出置信度高于设定阈值的人物图像框坐标信息，通过坐标信息可以提取出待测图片中行人图像集合$\boldsymbol{M}=\{\boldsymbol{M}_1,\boldsymbol{M}_2,\dots,\boldsymbol{M}_o\}$（其中$o$为原始图片中提取出人物图像的数量）；（2）将经过坐标信息裁剪后的行人图片作为经过训练后的网络的输入，网络最后输出的特征向量$\boldsymbol{m}=\{\boldsymbol{m}_{1p},\boldsymbol{m}_{2p},\dots,\boldsymbol{m}_{op}\}$（其中$p$为网络模型输出的特征向量的维度），同时将标准库中的所有图片集合$\boldsymbol{N}=\{\boldsymbol{N}_1,\boldsymbol{N}_2,\dots,\boldsymbol{N}_q\}$（其中$q$为网络模型输出的特征向量的维度），提取为与$\boldsymbol{m}_i$维度相同的特征向量$\boldsymbol{n}=\{\boldsymbol{n}_{1p},\boldsymbol{n}_{2p},\dots,\boldsymbol{n}_{qp}\}$，计算$\boldsymbol{m}$中任意一个特征$\boldsymbol{m}_{ip}$与$\boldsymbol{n}$中所有的特征向量的距离，并取距离最小值中的$k$值为$\boldsymbol{m}_i$的类标，即

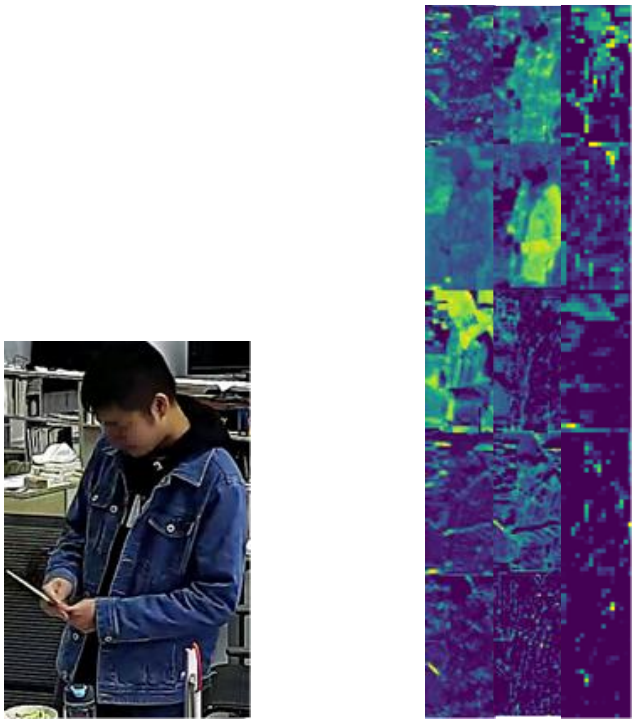
$$arg\min_k \beta\left(\sum_{j=1}^{p} m_{ij}^2+\sum_{j=1}^{p} n_{ij}^2\right)+\alpha\left(n_k * m_i^T\right) \tag{3-1}$$

式中，α，β为距离系数，一般设置为-2，1。$\boldsymbol{m}_i$表示为待测行人图片的特征，$\boldsymbol{n}_k$表示为标准库的图片特征，k表示为标准库特征对应的类标。

3.1.3 基于动态行人库的识别算法

目前 3.1.2 节采取的方法是基于表征学习的行人重识别方法，通过卷积神经网络从原始图像数据中根据需求自动提取表征特征，并验证一对行人图片是否属于同一个人。如图 3-2 所示，这是经过训练后的 Resnet50[38]的前 3 个 Layer 的输出的前五层特征图。由图可以看出经过训练过的神经网络模型往往是基于待

测图片的非生物特征如衣服颜色，样式，甚至背景信息等等来进行识别，这类特征往往会受到时间因素的干扰，因为长时间下来会存在行人衣着变换，视角变换等等情况。基于这种情况，我们提出动态行人库的方法，在识别过程中动态的更新标准库图片，使其尽可能的捕捉到行人特征变化的信息，以提升整个方法的识别率。



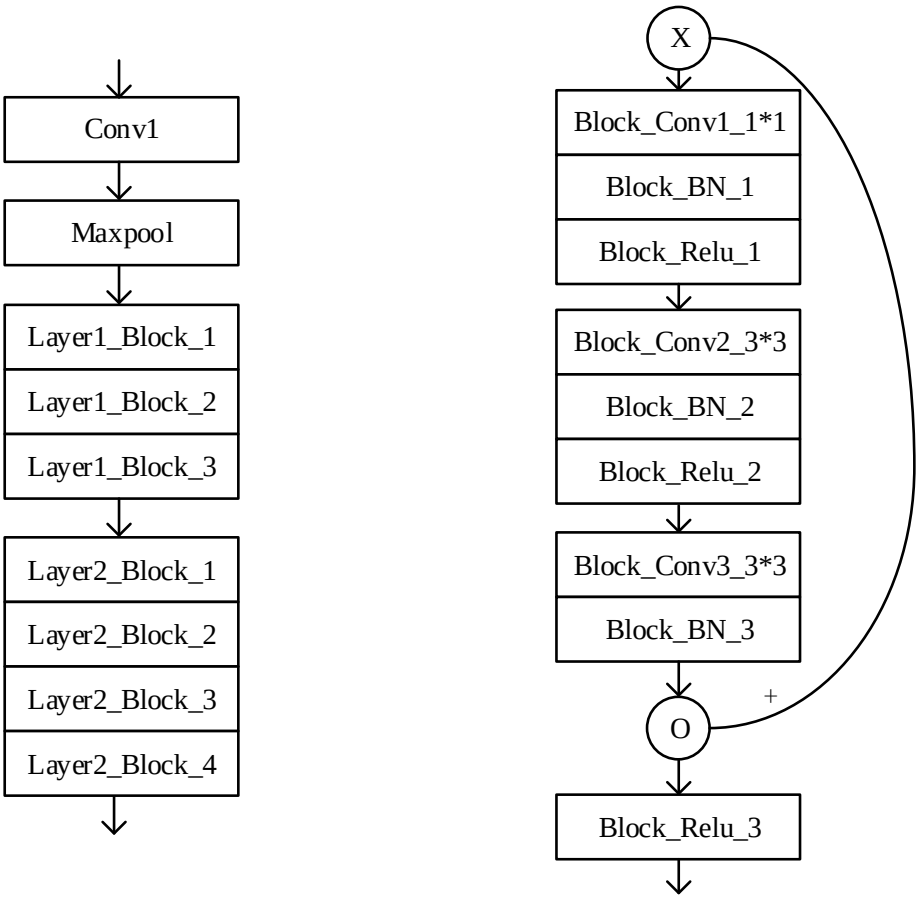
(a)原图 **(b)特征可视化图**
(a)Original image **(b)Feature visualization diagram**

图 3-2 特征敏感图
Figure 3-2. Feature sensitivity diagram

动态行人库方法的具体细节如下：将图片中每一张被目标检测算法检测出来的行人图片$\boldsymbol{M}$放入经过训练的新型卷积网络模型中提取出行人特征$\boldsymbol{m} = \{\boldsymbol{m}_{1p}, \boldsymbol{m}_{2p}, \dots, \boldsymbol{m}_{op}\}$，通过不断与标准库的特征$\boldsymbol{n} = \{\boldsymbol{n}_{1p}, \boldsymbol{n}_{2p}, \dots, \boldsymbol{n}_{qp}\}$（$p$ 表示为标准库中不同类标的个数）进行对比，如果最小距离达到一定阈值A时，我们将这张图片的特征$\boldsymbol{m}_{ip}$加入到标准库$\boldsymbol{n} = \{\boldsymbol{n}_{1p}, \boldsymbol{n}_{2p}, \dots, \boldsymbol{n}_{qp}, \boldsymbol{m}_{ip_h^r}\}$（其中$h$为最小距离的类标，$r$为当前类标的标签号）中辅助后续的识别。同时考虑到在实际场景下识别速度的要求，本文规定每个人新增的图片不得超过Q张，当存在超过的规定数目Q的行人特征$\boldsymbol{m}_{ip_h^{Q+1}}$时，则将$\boldsymbol{m}_{ip_h^1}$替换成$\boldsymbol{m}_{ip_h^{Q+1}}$继续进行识别，并依次修改标准特征库中的特征$\boldsymbol{m}_{ip_h^r}$的标签号$r$，使新增特征按照时间先后顺序排列。

3.2 网络训练及改造

3.2.1 网络改造



(a)网络结构图
(a)Network structure diagram

(b)Block结构图
(b)Block structure diagram

图 3-3 网络结构图以及 Block 图

Figure 3-3. Network structure diagram and block structure diagram

基于表征学习的卷积神经网络是针对公开数据集的数据特征及要求而进行训练的，因此基于公开数据集的卷积神经网络模型会存在跨域匹配的问题，本文通过对网络的改造来改善这一问题。本文的行人重识别网络采用了基于残差网络 Resnet50[38]的形式，结合 Yu 等人[39]提出中层特征具有特别表征信息的思想，实现监控视角下人物图片的特征提取。通过对残差网络中的中层网络输出进行改造，以此来提取含有不同层次的特征输出，使深度卷积网络模型在其他数据集中具有较好效果。

由于希望从行人图片中得到更多有效的特征，本文通过提取骨干网络的中层特征进行处理，将处理过的中层特征与最终特征进行拼接得出最终的输出。因为监控视角下的行人图片的表征特征更难提取，本文将 Resnet50 网络做如下的处理，保留了 Resnet50 网络结构的低层特征结构，保留的结构如图 3-3(a)所

示。其中 Block 表示残差网络特有的结构，Block 的结构如图 3-3(b)所示，共含有 3 个卷积层，前两个卷积层后紧接着 BN 层和 Relu 层，最后一个卷积层的输出先经过 BN 层，再与输入相加后经过 Relu 层。

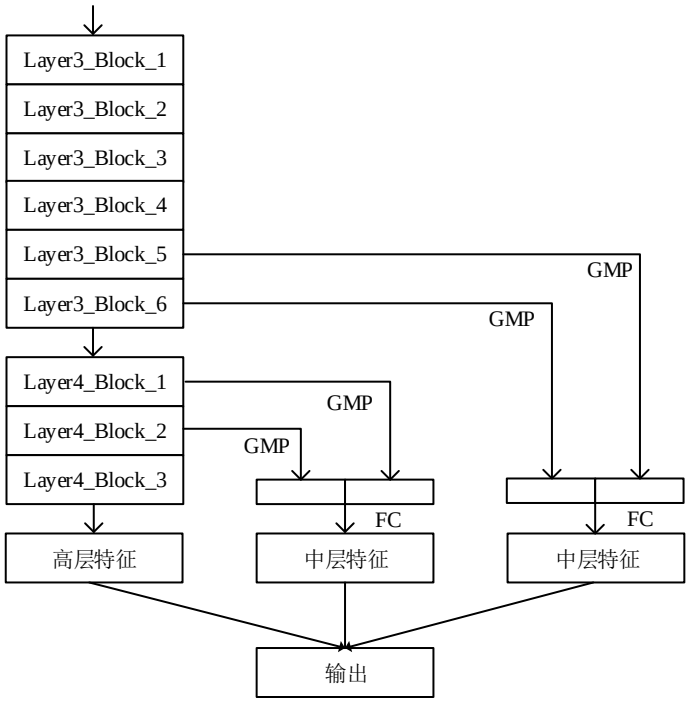


图 3-4 网络结修改图

Figure 3-4. Network modification diagram

深度卷积网络拥有自动提取特征并组合成高层特征的能力，所以基于表征学习的深度卷积网络往往侧重于高层特征。本文考虑利用深度卷积网络的中层特征，将中层特征通过池化策略与高层特征进行融合，得出具有中高层次的特征输出。进行改造的原因是注重纹理信息的中层特征会对具有抽象的高层特征存在解释性，互补之后的特征更具有辨别能力。因此为了更好的提取行人图片中的特征，本文采用了对网络中层特征进行处理的办法。如图 3-4 所示，将 Layer3_Block_5 ， Layer3_Block_6 ， Layer4_Block_1 ， Layer4_Block_2 和 Layer4_Block_3 的特征图先进行全局最大池化，之后将 layer3_block_5，layer3_block_6 分成一组，layer4_block_1，layer4_block_2 分成一组进行拼接，并加入全连接层进行降维操作，最后再和 layer4_block_3 进行拼接，得到最终的卷积神经网络输出。

3.2.2 网络训练

一个好的深度神经网络模型需要有大量的数据进行训练，数据量的大小直接影响网络的性能，使用少量的数据训练网络模型会造成网络产生过拟合现象，影响模型泛化能力。为了解决这个问题，目前传统的模型训练都是采用基

于迁移学习的方法，先进行预训练，之后再针对特定任务进行模型微调，借此达到要求效果。

对此本文采用基于迁移学习策略进行网络的训练。选择已经经过 ImageNet 数据集训练过的 Resnet50 网络模型，并依照 3.2.1 节的思路对其进行改造，最后对经过改造后的模型放入 Market1501 数据集中进行训练。

考虑传统训练卷积神经网络模型的方式，我们选择使用 Adam 优化算法对网络进行优化。Adam 优化算法结合 AdaGrad 和 RMSProp 两种优化算法的优点，具有实现简单，计算高效，对内存需求少等优点，在很多情况下算作默认工作性能比较优秀的优化器。Adam 公式如下：

$$m_t = \beta_1 \times m_{t-1} + (1-\beta_1) \times \mathrm{g}_t \tag{3-2}$$

$$v_t = \beta_2 \times v_{t-1} + (1-\beta_2) \times \mathrm{g}_t^2 \tag{3-3}$$

$$\widehat{m}_t = \frac{m_t}{(1-\beta_1^t)} \tag{3-4}$$

$$\hat{v}_t = \frac{v_t}{(1-\beta_2^t)} \tag{3-5}$$

$$\theta_t = \theta_{t-1} - lr * \frac{\widehat{m}_t}{(\sqrt{\hat{v}_t} + \varepsilon)} \tag{3-6}$$

式中，m_t，v_t分别为一阶动量项和二阶动量项，β_1，β_2为衰减率，$\widehat{m}_t$，$\hat{v}_t$为各自的修正值，lr为学习率，ε为极小值防止分母为 0。由表达式可以看出，Adam 优化算法可以从一阶及二阶动量进行调节，具有良好的效果。

其次在网络训练过程中，损失函数对训练效果的作用是非常重要的。损失函数用于衡量机器学习模型的预测能力。我们考虑将传统的交叉熵损失函数与行人重识别中的 TriHard Loss 损失函数同时用来对网络进行训练，训练公式如下：

$$L_{CE} = -\sum_i y_i \times \frac{\exp(x_i)}{\sum_j \exp(x_j)} \tag{3-7}$$

$$L_{HT} = \frac{1}{batchsize} \times \sum_{a \in \boldsymbol{batch}} \max(0, -1 \times (\min_{n \in B} d_{a.n} - \max_{p \in A} d_{a.p} + \gamma)) \tag{3-8}$$

$$L_A = \alpha \times L_{HT} + \beta \times L_{CE} \tag{3-9}$$

式中，x_i为模型预测结果，y_i为真实结果的分布，γ为人为设定的阈值参数，$\boldsymbol{Batch}$为一次训练所需图片的特征的集合，$batchsize$表示为一个$\boldsymbol{Batch}$的大小，a为单张行人图片的特征，A为$\boldsymbol{Batch}$中类标与a相同的行人图片特征集合，B为$\boldsymbol{Batch}$中类标与a不相同的行人图片特征集合。α，β分别为交叉熵损失

及 Tri-Hard 损失的系数，$d_{a.n}$为行人图片a的特征向量与负样本图片n的特征向量的欧式距离。

3.3 实验研究与分析

为了验证 3.2 节描述的方法，本文首先在服务器上搭建相关深度学习软件环境，并依据数据集做了人脸识别实验，对比实验等相关实验以验证算法有效性。

3.3.1 实验环境

本文所用方法使用服务器进行训练，其相关软硬件配置如下。

表 3-1 实验环境

Table 3-1. The experiment environment

参数名		参数值
处理器	型号	i7 7800X
	频率	3.5GHz
GPU	型号	GTX1080Ti
	容量	11G
	数量	2
内存	型号	DDR4 2666
	容量	16G
	数量	2
操作系统	型号	Ubuntu16.04
深度学习框架	类型	Pytorch0.4.0

3.3.2 数据集

Market1501 数据集是由 Zheng 等人[40]提出来的行人重识别数据集，整个数据集由 1501 个不同行人的图片组成，将其分成训练集和测试集，其中图片内容如图 3-5 所示。训练集中包含 751 个行人，共有 12936 张图片。测试集中包含 750 个行人，共有 19732 张图片。另外还有 query 文件，文件里为测试集中 750 个行人在每个摄像头中随机选择一张图片作为 query，最多每个人会有 6 张图片，共有 3368 张图片。这些数据都是由 6 个摄像头在清华大学中采集，并在 2015 年公开。

为了验证网络在实际场景下的泛化性能，本文基于室内办公场景下制作了室内的视频数据集，共包括 11 天的视频文件，室内共有 3 个不同视角的摄像头进行拍摄。考虑到验证算法有效性的问题，对视频文件做以下处理。对每个视频每隔 1 秒取 1 帧图片，总共有 10555 张监控视角下的原始图片。在将原始图片

放入 YOLOv3 目标检测算法中检测，共标出 10087 个目标，其中经过筛选共有有效目标 7610 个，共有 17 个实例，每个实例都有一张正脸照供人脸识别使用。另外这 11 天视频中，每天有单独的标准库，每个标准库里有出现在当天视频里人物的 2 张照片为卷积网络模型提供参照。

图 3-5 数据集实例图

Figure 3-5. Dataset instance diagram

3.3.3 实验数据分析

为了验证传统的人物识别方法在监控视角下的效果，我们仅仅简单的将人物的查全率作为指标。将本文数据集放入传统的人脸识别方法中，其查全率对比见表 3-2。

人脸识别方法采用是基于业内领先的 C++开源库 Dlib 的深度学习模型，在 LFW 数据集中高达 99.38%的准确率的模型和 Zhang 等人提出专门用于人脸检测的 MTCNN 网络。MTCNN 使用 3 个卷积网络以级联方式连接，将人脸检测和人脸特征点检测同时进行。

表 3-2 人脸查全率对比

Table 3-2. The comparison of face recall rate

方法	检测人脸个数	有效目标	参数值
Dlib	1098	7610	14.43%
MTCNN	1914	7610	25.15%

实验结果表明，在不考虑识别效果，仅仅考虑识别出人脸的情况下，模型的查全率连 30%都不能达到，这还是在没有考虑识别误差的情况下。因此在监

控视角下的人物识别方法里，人脸识别方法没有太多的效果。其可能原因有以下几点：

（1）监控视角下的视频信息的质量受很多因素的影响，如光照变化，拍摄目标距离远近等等，从而往往难以得到高质量的图片，不能满足目前人脸识别方法的要求。

（2）在监控视角下无法得到与现有人脸识别应用里识别的图片要求。目前人脸识别应用都是要求拍摄到人物的正脸照片，而监控视角下由于安装原因通常无法获得高质量的正脸图片，会导致人脸识别效果差。

（3）监控视角下人物活动具有随机性，会出现只有侧脸甚至背影的照片。这些照片无法获得人脸信息，自然也无法进行识别。

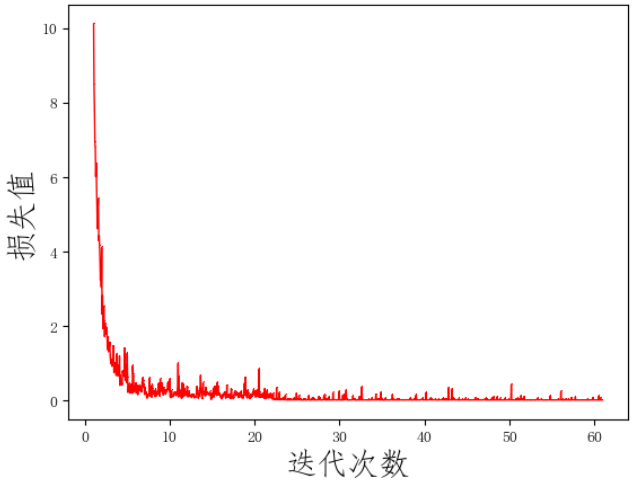


图 3-6 损失值迭代图

Figure 3-6. Iterative graph of loss values

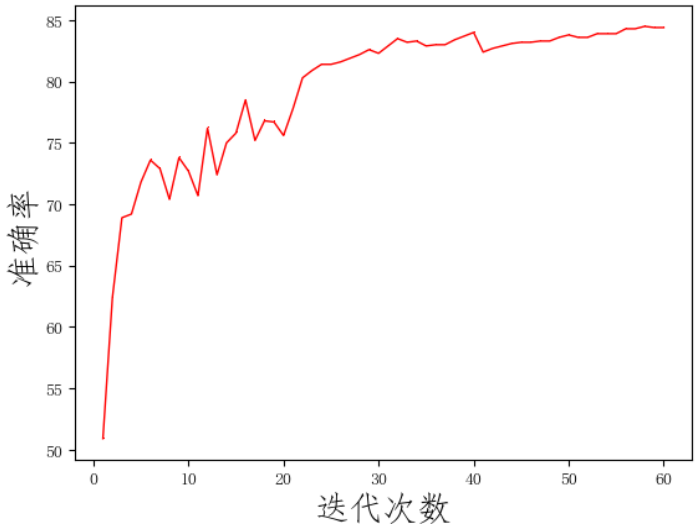


图 3-7 准确率迭代图

Figure 3-7. Iterative graph of accuracy iteration

为了验证训练模型在实际场景下的鲁棒性和泛化性，使用两种数据集下验证其效果，分别是 Market1501 以及自建数据集。模型使用 Adam 优化算法进行训练，训练过程中，损失函数值与识别率变化曲线如图 3-6 和图 3-7 所示。

参数配置为$\beta_1 = 0.9$， $\beta_2 = 0.999$，$lr = 0.003$，$\varepsilon = 1e-8$，训练数据集为 Market1501。对数据集图片进行以下的一些预处理，先经过随机 2D 裁剪，有几率将图片扩大至 1.125 倍，并随机裁剪一部分，最后将尺寸统一为256×128像素大小。接着进行随机水平翻转以及标准化的预处理步骤。将训练 Epoch 设置为 60，训练完成后保存最佳模型，此时在 Market1501 测试集上的识别率为 84.59%。针对多个不同的模型得到的识别率如表 3-3 所示。实验结果表明：

本文所设计的网络模型在训练数据集 Market1501 中验证效果比不上目前提出来的行人重识别网络模型，但是在公开数据集 Duke-MTMC 及自建数据集中所设计的网络模型相比于其他模型具有良好效果，说明在 Market1501 训练的模型对 Market1501 数据集上存在过拟合的现象，将模型用于其他数据集中准确率迅速下降。其次修改模型在其他数据集中效果好说明修改模型的泛化能力增加，并说明改造的有效性。

本文改造后的网络和 Resnet50m 网络相比，最大的改变是关于池化层的选择。这也证实了 Yu 等人提出对于不同形态的数据集需要使用不用的池化操作，达到跨域识别的特点，给以后新场景下的人物识别任务提供了新的思路。

表 3-3 不同模型识别率对比

Table 3-3. Recognition rates of different model

方法	Market-1501	Duke-MTMC	自建数据集
Resnet50	83.6%	19.1%	58.2%
Resnet50m	88.0%	20.6%	57.5%
MLFN[41]	87.8%	22.0%	61.8%
PCB[19]	86.5%	26.8%	55.4%
SENet[42]	87.8%	23.3%	62.1%
本文	84.6%	28.8%	66.0%

为了验证动态行人库的算法对网络识别的影响，分别对不同网络在识别过程中加入动态行人库算法，经过了多次尝试，对每个网络设置了最适合的阈值并做了几组对比试验，实验结果见表 3-4。由此可看出。

（1）动态行人库算法在一定程度上还是具有鲁棒性的，在所有数据集中都有不同程度的提升。这也验证了行人重识别对标准库的数量还是有依赖性，所以目前行人重识别在实用层面还需要考虑对标准库的改造。

（2）动态行人库算法的提升效果并不太大，这与初始想法有一定差异。目前存在的问题是当标准库图片放入了错误的照片后，有可能会一直被错误的照片影响识别准确率，这个是目前需要改进的地方。其次动态行人库的阈值的设定需要有一个具体的算法，考虑到网络结构的不同，阈值也要改变以适应网络。

表 3-4 不同模型准确率对比

Table 3-4. Accuracy rates of different model

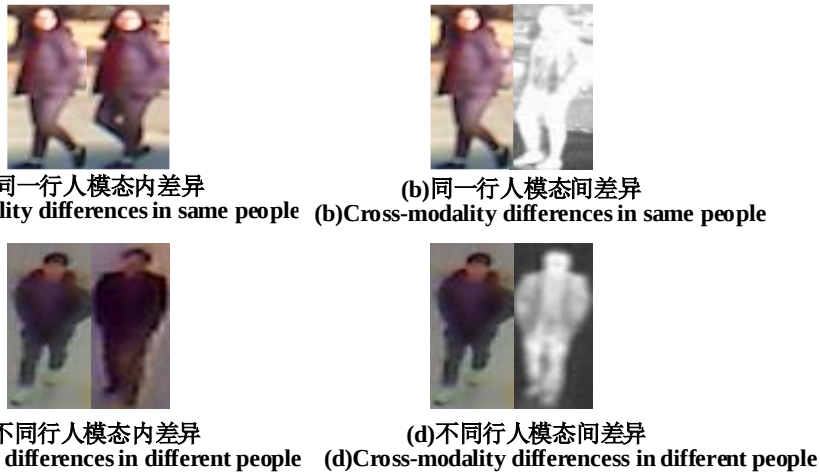
方法	不加入动态行人库算法	加入动态行人库算法
Resnet50	58.2%	59.8%
Resnet50m	57.5%	60.6%
MLFN[41]	61.8%	63.3%
PCB[19]	55.4%	57.6%
SENet[42]	62.1%	65.5%
本文	66.0%	69.3%

3.4 本章小结

针对监控视角下人物识别的问题，本文提出了一种基于深度卷积网络的方法，首先针对实际场景下选择合适的网络结构对骨干网络进行改造，增加其跨域识别的能力。考虑实际场景下数据集不够的问题，采用基于迁移学习的方法，在公用数据集上进行训练，并增加数据预处理等方式加强训练效果。考虑到行人重识别算法对于时效性的要求，提出动态行人库算法，增加模型的识别率，这对于实际场景中识别算法的使用和模型结构的改造和训练具有一定的启发意义。

第四章 基于可见光-热成像模态行人重识别方法研究

由图 4-1 所示，跨模态行人重识别领域内不仅存在模态间差异问题，还存在传统行人重识别领域内姿态、视角等差异性问题[43-47]。本文提出多种改进方法提升跨模态行人重识别模型识别精度，并在公开数据集 RegDB 数据集上进行实验验证方法的有效性。本文基于可见光-热成像跨模态数据集中存在多种模态的特点，使用双流结构和改进的特征提取模块分别提取可见光模态和热成像模态图片的特征信息，其次对传统困难三元组损失函数（Hard Triplet Loss）[37]进行改进，提出跨模态双流困难三元组损失进行模型的训练，最后为了提升模型训练中对困难样本学习的权重，本文使用焦点损失函数替代传统交叉熵损失函数提升模型训练效率。利用上述方法训练的模型在 RegDB 数据集实验中的各项指标均有提升。



(a)同一行人模态内差异
(a)Intra-modality differences in same people

(b)同一行人模态间差异
(b)Cross-modality differences in same people

(c)不同行人模态内差异
(c)Intra-modality differences in different people

(d)不同行人模态间差异
(d)Cross-modality differencess in different people

图 4-1 跨模态行人图片变化示意图

Figure 4-1. Schematic diagram of cross-modality people picture changes

4.1 多重改造方法

本文提出方法整体框架如图 4-2 所示。整体框架包括深度卷积网络特征提取模块以及特征训练模块。本文基于跨模态行人重识别中双模态的特点，使用双流结构分别提取可见光模态行人图片特征以及热成像模态行人图片特征。其次，对特征提取模块进行改造，提升模型提取特征能力。特征提取完成后，分别经过排序损失以及身份损失训练，确保卷积神经网络模型能学习辨别性特征。

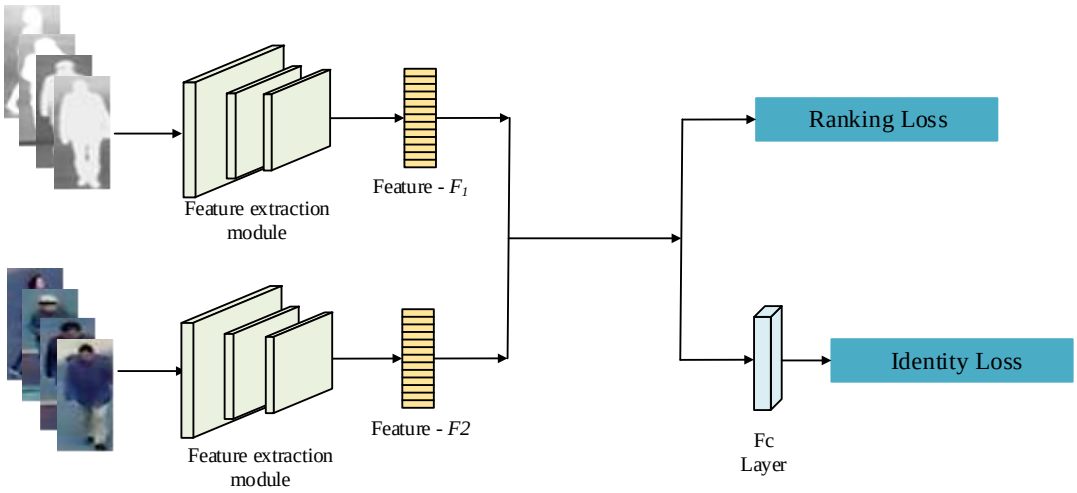


图 4-2 框架结构图

Figure 4-2. Framework structure figure

4.1.1 特征提取模块改造

深度卷积网络模型强大的特征提取能力是其能受到广泛应用原因，但是由于卷积网络模型在训练过程中存在对数据过拟合的问题，限制了卷积网络模型的表现。本文基于可见光模态研究中的中层特征扩展方法的优异表现，将该方法应用在跨模态行人重识别模型的特征提取模块中。

本文选取 Resnet50 网络作为骨干网络，作为特征提取模态的基础。Yu 等人对 Resnet50 的 Layer_4 层进行改造，将 Layer4 中 Block_1 和 Block_2 卷积块的特征提取后进行池化、拼接及全连接层等操作，组成 1024 维的中层特征。中层特征和 Layer4 中的 Block_3 卷积块输出特征进行拼接，组成 3072 维的特征向量，这其中包含大量的中层特征，有助于增加卷积网络对行人图片的特征表达。其中 Layer 和 Block 的相关资料可参考图 3-3。

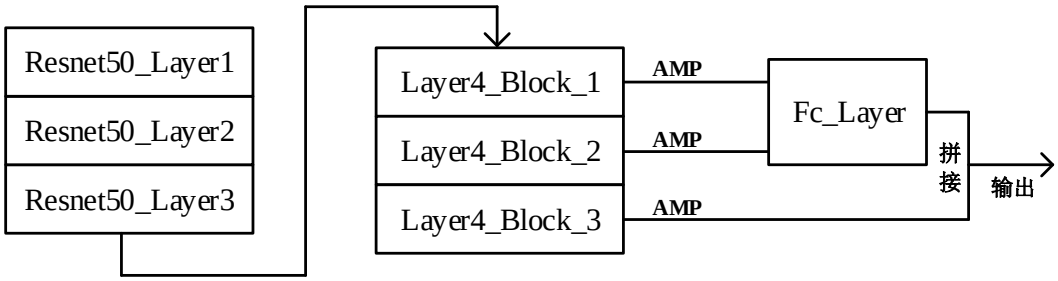


图 4-3 Resnet50m 结构图

Figure 4-3. Resnet50m structure figure

为了进一步提升卷积网络模型的泛化能力，本文于第三章提出中层特征扩展方法，将 Layer3 中的 Block_5，Block_6 特征块的输出特征，与 Layer4 中 Block_1 和 Block_2 特征块的输出特征进行分组操作。其中 Block_5 和 Block_6 的输出特征，Block_1 和 Block_2 的输出特征经过池化全连接层处理分别输出为

512 维的特征，最终和 Block_3 的输出特征组成 3072 维的特征向量。Yu 等人提出的方法结构如图 4-3 所示。

4.1.2 排序损失

排序损失（Ranking Loss）广泛运用于人脸识别，行人重识别领域，通常用于降低类内距离。经典的排序损失函数为三元组损失函数（Triplet Loss）[33]，它不仅具有降低类内距离的特点，还有增加类间距离的能力。为了更好地提升三元组损失函数效果，Hermans 等人[37]提出困难三元组损失函数，通过对样本进行困难采样的方式，提升模型训练效率。困难三元组损失每次从训练集中提取P个行人，每个行人抽取K张照片，总共PK张照片构成一个$\boldsymbol{Batch}$。困难三元组损失函数公式如(4-1)所示：

$$L_{Htri}=\frac{1}{P\times K}\sum_{i=1}^{p}\sum_{a=1}^{K}[\max_{p=1,\dots,K}D\left(f\left(x_a^i\right),f\left(x_p^i\right)\right)-\min_{\substack{j=1,\dots,P\\ n=1,\dots,K\\ j\neq i}}D\left(f\left(x_a^i\right),f\left(x_n^i\right)\right)+\alpha]_+ \tag{4-1}$$

$$D\left(f\left(x_a^i\right),f\left(x_p^i\right)\right)=\left\|f\left(x_a^i\right)-f\left(x_p^i\right)\right\|_2 \tag{4-2}$$

其中x_a^i表示当前图片为锚点，x_p^i和 x_n^i分别表示为与x_a^i类标相同的行人图片和类标不同的行人图片。$f\left(x_a^i\right)$为卷积网络提取获得行人图片的特征向量表示，α为人为设定阈值参数，当x_a^i与x_n^i之间欧式距离和x_a^i与x_p^i之间欧式距离之差小于α时则$[b]_+=b$，当大于α时，则$[b]_+=0$，其中$[b]_+=\max(b,0)$。

困难三元组因为在训练中需要对$\boldsymbol{Batch}$里所有行人图片进行遍历，每一张锚点行人图片需要在$\boldsymbol{Batch}$中寻找一个与锚点行人图片相比最难（距离最远）的类标相同的行人图片和最难（距离最近）的类标不同的行人图片组成一个困难三元组，所以共有PK个困难三元组。由于困难三元组损失每次训练时都专注于采样$\boldsymbol{Batch}$中的难样本，所以在训练速度以及模型精度上优于三元组损失函数。

$$L_{Dtri}=\frac{1}{2\times P\times K}\sum_{i=1}^{p}\sum_{a=1}^{2K}[\max_{p=1,\dots,2K}D\left(f\left(x_a^i\right),f\left(x_p^i\right)\right)-\min_{\substack{j=1,\dots,P\\ n=1,\dots,2K\\ j\neq i}}D\left(f\left(x_a^i\right),f\left(x_n^i\right)\right)+\alpha]_+ \tag{4-3}$$

为了将困难三元组损失用于跨模态行人重识别问题中，本文对困难三元组损失进行改造，称为跨模态双流困难三元组损失。该损失将困难样本采样范围扩大为两个可见光模态和热成像模态中，采样数量为传统困难三元组的两倍，

每个$\boldsymbol{Batch}$中含有P个行人，每个行人含有K张可将光模态的行人图片，K张热成像模态的行人图片，一共有$2PK$张行人图片。跨模态双流困难三元组损失如公式(4-3)所示。其中锚点图片x_a^i选择范围是可见光模态和热成像模态的行人图片的并集合，x_p^i为与x_a^i类标相同的同模态行人图片或跨模态行人图片，x_n^i为与x_a^i类标不同的同模态行人图片或跨模态行人图片，其余符号与公式(4-1)和公式(4-2)相同。

其次本文为了更好的提升跨模态行人重识别模型的识别效率，本文还使用了 Zhu 等人[48]提出的异型中心损失。异型中心损失提出使用两个模态行人图片分布的中心距离作为相似性判断依据，而不是两个模态行人图片的分布之间的距离。异型中心损失的公式如下：

$$L_{HC} = \sum_{i=1}^{U} [\left\| \boldsymbol{c}_{i,1} - \boldsymbol{c}_{i,2} \right\|_2^2] \tag{4-4}$$

$$\boldsymbol{c}_{i,1} = \frac{1}{M} \sum_{j=1}^{M} \boldsymbol{x}_{i,1,j} \tag{4-5}$$

其中$\boldsymbol{c}_{i,1}$和$\boldsymbol{c}_{i,2}$分别表示可见光模态和热成像模态的第i个类标的行人图片的中心特征分布，M表示可见光模态的行人图片的数量，$\boldsymbol{x}_{i,1,j}$表示第j个可将光模态的第i个类标的行人图片。

本文分别对以上两种损失在 RegDB 数据集上进行了实验，实验结果如 4.2.4 节所示。

4.1.3 身份损失

身份损失是深度卷积网络训练中基本的损失函数，用于减少类内之间距离。常用的身份损失为交叉熵损失，其广泛运用在图像分类，图像检索等领域。然而，在可见光-热成像跨模态行人重识别领域内，由于热成像技术无法提取颜色，纹理等信息，使用热成像技术获取的行人图片缺少具有辨别性的有效特征，只有行人的轮廓特征，这导致可见光-热成像跨模态行人重识别模型识别热成像模态的行人图片十分困难。由于数据集中出现难以识别的样本，卷积网络的训练应该多注重难样本行人图片，但是交叉熵损失没有区别对待训练样本中难易样本学习的权重，所以交叉熵损失函数无法获得较好效果。

为了针对可见光-热成像跨模态行人重识别中出现的难易样本的问题，本文使用 Lin 等人[26]提出的焦点损失函数处理难易样本问题。焦点损失函数不同于传统处理类别失衡算法忽略易分类样本，它只是降低易分类样本的权重，同时提升难样本学习的权重。焦点损失函数公式和交叉熵损失函数如下：

$$L_{CE} = -\sum_{i=1}^{M} y_i \, log \, \hat{p}_i \tag{4-6}$$

$$L_{Fo} = \frac{1}{2 \times P \times K} \sum_{i=1}^{2PK} \sum_{j=1}^{N} -(1 - p_{ij})^{\gamma} log(p_{ij}) \tag{4-7}$$

其中，y_i为类别真实概率分布，$\hat{p}_i$表示经过 softmax 函数计算的x_i类别预测概率，M为行人类别个数，其中$(1 - p_{ij})$为交叉熵损失函数调节因子，p_{ij}表示$\boldsymbol{Batch}$中第i张照片第j个类标对应概率，N表示验证集类标总数，γ是可调参数，本文统一设置值为 2。实验结果如 4.2.5 节所示。

4.2 实验结果与分析

4.2.1 数据集描述及评价指标

本文使用公开的可见光-热成像跨模态行人重识别数据集 RegDB 评估上述章节提出的方法。RegDB 数据集是由 Nguyen 等人[23]收集并公开的具有权威性的跨模态行人重识别数据集，内容包括在同一时刻下通过可见光相机和热成像相机拍摄人物图片信息。数据集一共有 412 个行人类别，每个行人类别都有 10 张可见光图片以及 10 张热成像图片。数据集中包含正面拍摄以及背面拍摄。由于拍摄期间人物会进行无规则运动，每张相片都存在姿态，光线条件差异。数据集共有 4120 张可见光图片以及 4120 张热成像图片。

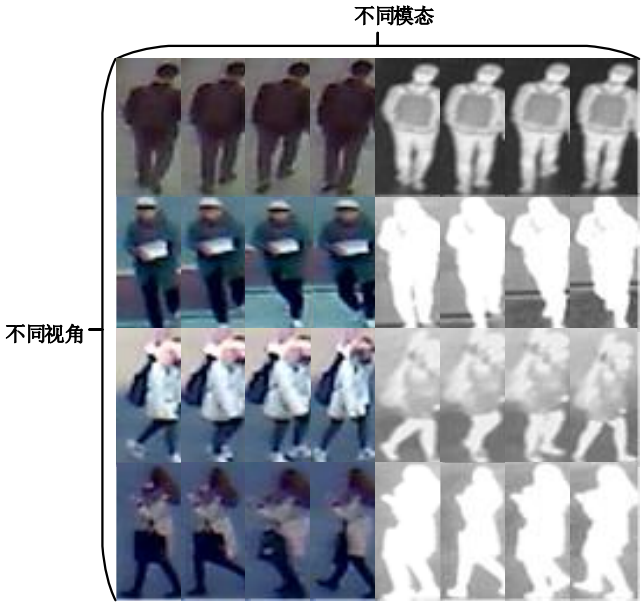


图 4-4 RegDB 数据集示例图

Figure 4-4. RegDb dataset example graph

Ye 等人[46]针对 RegDB 数据集的特点，提出一套公开验证方法用于验证可见光-热成像跨模态行人重识别模型的效果。其中，Ye 等人对 RegDB 数据集使用 10 折交叉验证方式，每次从 412 个行人中选择 206 个行人共 4120 张作为训

练集，其余 4120 张图片用于验证。重复进行 10 次实验并对总的实验数据取平均平均值作为单个实验最终指标，RegDB 数据集内行人图片如图 4-4 所示。

在实验验证阶段，Ye 等人使用平均精度均值（mAP）以及累计匹配曲线（CMC 曲线）作为评价指标。mAP 常用于衡量多标签图像分类的效果，是一种常见的指标。CMC 曲线用于评估行人重识别算法算法性能，通常评价指标包括 rank1，rank10 等。mAP 计算公式如下：

$$mAP = \frac{1}{N}\frac{1}{K_{q_i}}\sum_{i=1}^{N}\sum_{j=1}^{K_{q_i}}(\hat{r}_j/r_j) \tag{4-8}$$

其中N为需要检索图片的个数，K_{q_i}为需要检索图片q_i的 ID 在验证集中出现个数，其中验证图片集合记为G，图片q_i提取特征后依次和G中行人图片特征计算距离，并按照距离的升序排列，距离小的在前，并记为$\hat{G}$，将$\hat{G}$中与q_i相同 ID 的图片按照排列并另记为$\hat{G}_{q_i}$，$\hat{r}_i$表示$\hat{G}_{q_i}$中某一张 ID 与q_i相同图片$\hat{g}_i$在$\hat{G}_{q_i}$的位置，r_j表示为$\hat{G}$中$\hat{g}_i$在$\hat{G}$的位置。Rank-n 表示基于$\hat{G}$中前n个搜索结果中包含正确样本概率。

本文采用服务器训练跨模态行人重识别模型，显卡为 NVIDIA GeForce 1080Ti，显存为 11GB，操作系统是 Ubuntu16.04。算法使用 Pytorch 实现。

在实验参数设置中，每一次训练中随机选择P个行人类标，每个行人类标随机选择K张可见光模态行人图片，K张热成像模态行人图片，一个$\boldsymbol{Batch}$选择$2PK$张图片。其中本文设定P等于 8，K等于 4，一个$\boldsymbol{Batch}$中共有 64 张图片。在训练阶段本文对行人图片进行预处理，包括随机水平翻转，调整图片大小等。训练 Epoch 为 60，每一个结果需要进行十折交叉验证实验，最后取 10 个实验平均值作为结果，保证实验可靠性。

4.2.2 对比其他跨模态行人重识别方法

表 4-1 RegDB 数据集对比结果

Table 4-1. Comparison results of RegDB dataset

方法	Rank-1/%	Rank-10/%	Rank-20/%	mAP/%
GSM[49]	17.28	34.47	45.26	15.06
MLAPG[50]	17.82	40.29	49.73	18.03
XQDA[51]	21.94	45.05	55.73	21.80
HCML[46]	24.44	47.53	56.78	20.80
BDTR[21]	33.47	58.42	67.52	31.83
Ours	**37.62**	**64.17**	**75.73**	**40.85**

本文选取多个已经在 RegDB 数据集评估有效性的跨模态行人重识别算法。为了评估本文方法性能，实验评估指标采用 mAP，以及 CMC 曲线中 rank-1，rank-10 以及 rank-20。

在表 4-1 中，本文选择 Resnet50 作为骨干网络，使用双流结构分别提取可将光模态行人图片和热成像模态行人图片，运用中层特征扩展方法改造特征提取模块，并结合跨模态双流困难三元组损失和焦点损失进行训练，实验结果表明，本文提出方法在 RegDB 数据集上存在一定提升，并且在评价指标上要优于现有方法。

4.2.3 特征提取模块改造方法评估

本文选择resnet50作为骨干网络，并提出中层特征扩展方法改造特征提取模块。为了验证提出方法的效果，本文对 3 种特征提取模块分别进行实验，特征提取模块分别为 Resnet50，Resnet50m 以及 Resnet50m_ex 三种模型，Resnet50m_ex 是采用中层特征扩展方法改造的模型。其中分别使用相同的损失函数对特征提取模块进行训练，并使用 RegDB 数据集进行验证，实验结果如表 4-2 所示。

表 4-2 改造方法对比结果

Table 4-2. Comparison results of retrofit methods

损失函数	模型	Rank-1/%	Rank-10/%	Rank-20/%	mAP/%
L_{DTRI}	Resent50	8.20	19.32	28.79	12.29
	Resent50m	3.93	11.50	18.40	7.60
	Resent50m_ex	**25.39**	**44.66**	**55.68**	**30.17**
$L_{HC}+L_{CE}$	Resent50	25.10	51.99	63.98	20.95
	Resent50m	27.62	61.36	73.40	24.87
	Resent50m_ex	**32.43**	**65.00**	**76.89**	**30.18**

实验结果表明，本文提出中层特征扩展方法在跨模态行人重识别中能有效提升模型的识别准确率。相比 resnet50 以及 resnet50m 卷积网络具有稳定性以及准确率。

4.2.4 排序损失评估

本文对传统困难三元组损失进行改造，将其困难样本的采样范围扩大至可将光模态和热成像模态，并将检索图片增加为$2PK$张。为了验证跨模态双流困难三元组损失的有效性，本文将其和异型中心损失进行对比，采用不同损失函数的组合对模型进行训练，并在 RegDB 数据集上进行实验，实验结果如表 4-3 所示。

表 4-3 损失函数对比结果

Table 4-3. Comparison results of loss function

损失函数	Rank-1/%	Rank-10/%	Rank-20/%	mAP/%
L_{DTRI}	25.39	44.66	55.68	30.17
L_{HC}	1.02	3.20	5.34	1.62
L_{HC} +L_{CE}	32.43	65.00	76.89	30.18
L_{DTRI} +L_{CE}	**36.21**	**60.19**	**74.37**	**39.26**

由表 4-3 可以看出，单独使用排序损失训练跨模态行人重识别模型时，存在效果不明显的问题，跨模态双流困难三元组损失受到的影响会小于异型中心损失。但是当排序损失和身份损失共同使用时，效果提升较明显，说明跨模态行人重识别网络模型训练中，使用双重损失训练效果更好。

4.2.5 焦点损失评估

焦点损失用于缓解热成像模态行人图片中出现无法获取辨别性特征情况导致的样本失衡问题。本文将焦点损失用于跨模态行人重识别中，替代传统交叉熵损失训练模型。为了验证焦点损失函数的有效性，本文基于 4.2.3 和 4.2.4 的方法，仅仅通过替换交叉熵损失函数的方式，进行焦点损失评估。实验结果如表 4-4 所示。

表 4-4 身份损失对比结果

Table 4-4. Comparison results of identity loss

模型	损失函数	Rank-1/%	Rank-10/%	Rank-20/%	mAP/%
Resnet50	L_{DTRI}+L_{CE}	8.83	15.87	22.38	12.97
	L_{DTRI}+L_{Foc}	**9.81**	**17.43**	**24.71**	**13.64**
Resnet50m	L_{DTRI}+L_{CE}	9.61	17.57	24.27	12.80
	L_{DTRI}+L_{Foc}	**10.25**	**19.32**	**24.90**	**13.13**
Resnet50_ex	L_{DTRI}+L_{CE}	36.21	60.19	74.37	39.26
	L_{DTRI}+L_{Foc}	**37.62**	**64.17**	**75.73**	**40.85**

由表 4-4 可以看出，焦点损失在三个特征提取模型的实验中都比交叉熵损失表现好，特别当使用中层特征扩展方法的特征提取模块以及使用跨模态双流困难三元组损失作为排序损失时，跨模态行人重识别模型达到本文实验最好效果，并且在 RegDB 数据集上好于大多数跨模态行人重识别算法效果。

焦点损失能提升模型效果的原因在于可见光-热成像跨模态行人重识别数据集中存在类别失衡问题，由于热成像模态行人图片包含有效特征较少，属于难样本，在训练时模型没有注重难样本的学习，导致准确率无法提升。实验结果

证明，在 4.2.3 节和 4.2.4 节改进基础上，焦点损失还能提升跨模态行人重识别模型提取特征能力。

4.3 本章小结

针对跨模态行人重识别问题，本文提出多种改进方法提升跨模态模型识别精度。考虑可见光-热成像跨模态行人重识别中双模态的特点，使用双流结构分别提取可见光模态和热成像模态的行人图片，并且对特征提取模块进行改造。同时将传统单模态困难三元组损失改造成跨模态双流困难三元组损失，提升模型训练效果。基于跨模态行人重识别中类别失衡问题，本文使用焦点损失替换传统交叉熵损失函数，结合模型改造方法和损失函数的改进，在公开数据集 RegDB 上取得较好效果。

第五章 监控视角下人物识别系统搭建

基于第三章讲述的深度学习行人重识别算法以及目标检测算法，本文在室内环境下搭建了一套监控视角下人物识别系统，一方面用于制作室内环境下可将光模态行人数据集，用于验证经过表征学习的训练的卷积神经网络模型在实际场景下的效果，另一方面为视频监控网络下场景下实时进行行人识别提供初步方案。本章节将会阐述监控视角下人物识别系统的构建，其中包括室内环境参数选择，摄像头选型，人物识别系统识别流程等。

5.1 实验场景参数选择

监控视角下人物识别系统的实验环境是设计人物识别系统的基本，只有当确定实验场景后，才能依据场景内的参数选择合适的摄像头以及识别流程。场景内参数包括以下室内场景面积大小，室内环境物摆放两方面。本系统构建的需求有三个方面：（1）为了构建出一个具备可靠公信力的可见光行人重识别数据集，系统需要存在多个摄像头同时进行拍摄工作；（2）为了体现行人重识别算法能力，行人目标需要存在遮挡情况，增加行人重识别算法识别难度；（3）为了体现人物识别系统实时识别行人目标，需要控制实验环境大小保证在有限处理速度情况下尽可能将视角覆盖整个实验环境。

基于以上几种需求，监控视角下人物识别系统的室内场景面积大小决定在在$20M \times 8M$的室内空间中，其中这个空间中存在多个遮挡物，保证行人目标在室内移动时会出现收到遮挡情况。整体规划图和实际效果图如图 5-1，5-2 所示。

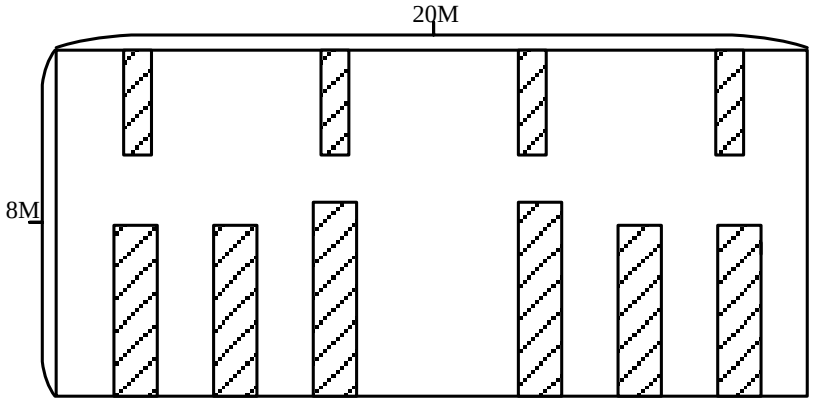


图 5-1 总体规划图

Figure 5-1. Overall plan figure

图 5-2 实际场景图

Figure 5-2. Actual scene figure

其次为了保证摄像头视角整体覆盖以及实时性处理要求，我们选择使用 3 个摄像头进行拍摄，三个摄像头位置如图 5-3 所示，分别置于位置 A，位置 B，以及位置 C 处。每个摄像头拍摄的范围如虚线所示，保证在使用少量摄像头的情况下能完成摄像头视角在室内场景里的全覆盖。

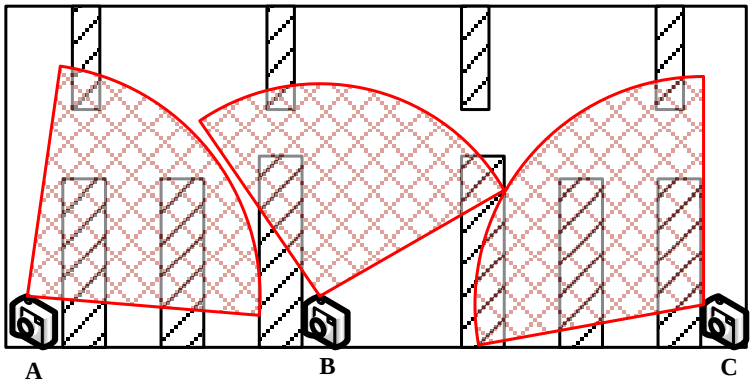


图 5-3 摄像头拍摄范围图

Figure 5-3. Camera shooting range figure

为了保证摄像头拍摄信息能实时读取并进行处理，需要建立摄像头和处理终端的内部局域网网络。本文使用局域网交换机进行网络摄像头与处理终端交换。将摄像头网络接口和处理终端网络接口以有线电缆插入，并将另一端接口插入交换机接口中，让处理终端能够获取摄像头拍摄数据。交换机图片及参数如图 5-4 及表 5-1 所示。

图 5-4 S1724G-AC 交换机图

Figure 5-4. S1724G-AC switch figure

表 5-1 S1724G-AC 交换机参数

Table 5-1. S1724G-AC switch parameters

参数名	数值	单位
包转发率	36	Mpps
交换容量	48	Gbps
固定端口	24	个
端口容量	10/100/1000Base-T	兆

5.2 摄像头参数选择

摄像头是监控视角下人物识别系统的核心，正确选择摄像头参数是获取高质量行人图片的重要步骤。摄像头的参数包括以下几个方面：（1）拍摄图片大小；（2）镜头焦距；（3）接口功能等三个方面进行选择。

首先对于拍摄图片大小，本文中采用图片大小标准为1920×1080，其中帧率为$50\text{Hz}: 25fps$。拍摄图片大小采用为 1920 * 1080 的依据在于在此分辨率下行人目标的长宽会在 200 - 300 区间内，这对于行人重识别网络来说是一个合适的输入大小。其次对于镜头焦距大小选择，由于室内拍摄环境大小为$20\text{M} \times 8\text{M}$，为了保证摄像镜头全覆盖，本文选择镜头焦距大小为 4mm，其拍摄角度为 69°，可拍摄的最远距离为 10M。

由于需要处理终端需要实时获取摄像头拍摄信息，所以摄像头需要存在通讯接口以及支持通讯协议。考虑到本文存在多个摄像头同时获取数据以及存储视频情况，所以使用 RTSP 协议保证视频流传输及存储。RTSP 式实时流传输协议，该协议指明一对多应用程序如何有效通过 IP 网络传送数据。RTSP 协议允许多个串流同一时间内进行需求控制，在同一时间内还能降低服务端网络用量。所以本文选择 DS-2CD3T20FD-I3W 型号摄像头，支持 RTSP 协议并附有 1 个 RJ45 10M/100M 自适应以太网口。其图片及参数列表如下。

图 5-5 DS-2CD3T20FD-I3W 摄像头

Figure 5-5. DS-2CD3T20FD-I3W camera

表 5-2 DS-2CD3T20FD 参数

Table 5-2. DS-2CD3T20FD parameters

参数名	数值	单位
分辨率	1920*1080	mm
镜头焦距	4	mm
支持协议	RTSP	个
以太网接口	1	个
接口容量	10/100	兆

5.3 监控视角下人物识别系统流程图

在系统硬件项目搭建完成后，需要设计监控视角下人物识别系统识别流程。监控视角下人物识别系统包含两个方面的任务，一方面是目标检测，另一方面是行人识别。其次，由于使用多摄像头网络进行拍摄，在处理终端资源空闲情况下，可以开启多线程同时处理多个摄像头视频数据达到实时处理效果。

5.3.1 人物识别流程图

监控视角下人物识别系统在识别前需要有一些准备工作，包括训练及行人重识别神经网络模型以及建立目标人脸库两方面。首先，行人重识别神经网络模型可以选择通用模型，也可以选择改进模型，本文使用基于残差网络改进的 ResNet50m 网络模型进行训练。其中先用 ImageNet 图像分类数据集进行预训练，训练完成后将训练后的卷积神经网络模型进行改进，去掉全连接层，增加中层特征融合操作，再使用 Market1501 数据集进行重新训练。其次，目标人脸库是人物识别系统的重要依据，需要将行人目标的人脸进行保存，在人脸识别的保证正确的基础上进行行人重识别。准备工作结束后，人物识别系统识别流程图如图 5-6 所示，分成 5 步依次进行说明。

人物识别步骤 1：行人检测，本文使用的行人检测算法为 YOLOv3 算法，首先需要加载预训练好的 YOLOv3 目标检测网络模型，其次将使用 RTSP 协议从摄像头读取到一帧图片送入 YOLOv3 目标检测网络中。根据网络模型输出的矩形坐标轴信息列表，从图片中裁剪出每个行人图片。

人物识别步骤 2：人脸识别，本文使用的人脸识别算法是 Face-Recognition 库中训练好的人脸识别模块，首先使用人脸识别模块的特征提取函数获取标准人脸库中的特征向量。其次将由步骤 1 中裁剪出来的行人图片作为人脸识别模块的检测人脸函数输入，若检测到行人图片中存在人脸，则从裁剪出来的行人图片中提取人脸特征并与标准人脸库中的特征向量进行对比，若相似度超过设

定阈值，则判断为人脸识别成功，输出身份信息，并将当前人物的行人特征和类标保存至标准行人库中。若检测到不存在人脸，则进行下一个识别步骤。

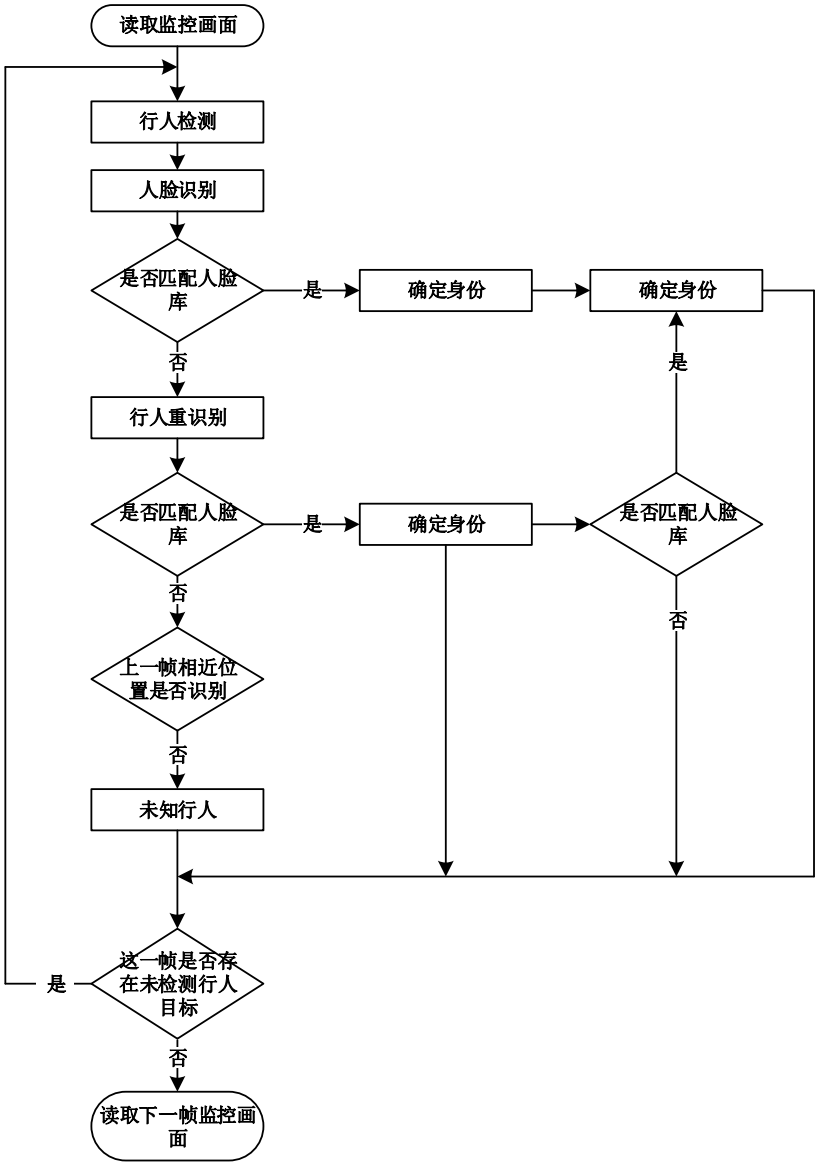


图 5-6 人物识别流程图

Figure 5-6. People recognition flowchat

人物识别步骤3：行人重识别，当人脸检测函数无法从行人图片中检测到人脸时，则使用预训练的行人重识别卷积网络模型提取行人特征，将提取到的特征和标准行人库中的行人特征进行相似度对比。若相似度大于设定阈值时，则判断行人重识别成功，输出身份信息，同时若相似度很高时，将提取到的行人特征保存至标准行人库中供下次识别使用。若相似度小于设定阈值或标准行人库中无行人特征时，则无法识别，进行下一个识别步骤。

人物识别步骤4：基于上下帧位置关系识别，当人脸检测和行人重识别都无法识别身份时，则通过上下帧行人之间的位置关系进行判断。首先将上一帧所有行人坐标保存作为下一帧照片寻找的依据，当下一帧中某个无法识别的行人

图片的位置区域与上一帧的某个确认行人目标区域重合度很高，则认定这个无法识别的行人图片的类标为上一帧确认行人的类标。若还是无法识别，则跳过此人进行下一个行人目标检测。

5.3.2 多线程识别流程图

由于人物识别系统有实时性需求，本文在基于处理终端资源空闲的情况下，考虑使用多线程进行同时识别，其中每条线程中运行 5.3.1 章节中人物识别流程。当多条线程同时处理完一个图片后，将其进行拼接并输出。

多线程识别流程中需要使用第三方环境进行使用，监控视角下人物识别系统是基于 Python 进行编写，多线程库是使用 Python 自带的 MultiProcess 库，读取摄像头图片的功能是由 OpenCV 库提供。多线程识别流程如图 5-7 所示。

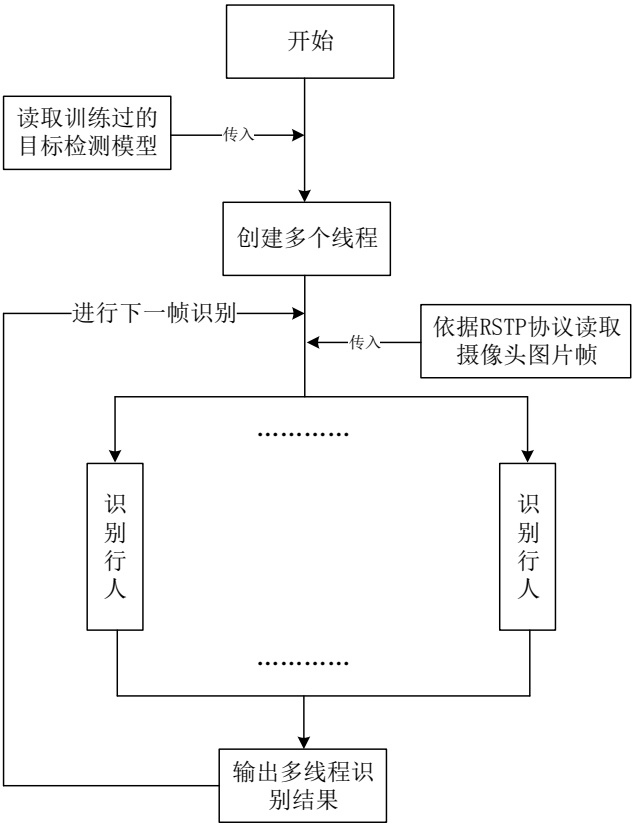


图 5-7 多线程识别流程图

Figure 5-7. Multi-thread recognition flowchat

多线程识别步骤 1：构建全局人脸库及行人库，作为线程内识别依据。其次读取目标检测网络模型以及行人重识别神经网络模型，作为全局变量，多线程重复读取模型时间。使用 MultiProcess 库依据处理终端资源配置创建合适的线程数。

多线程识别步骤 2：在每个线程内使用 OpenCV 库内函数依据 RSTP 协议读取摄像头视频，其中为了保证不会因为读取过快造成运行资源过大，每次读取

一帧，当线程中当前帧未识别完成时，后续帧不存储，直到线程处理完成后，才读取新的帧继续处理。

多线程识别步骤 3：读取帧图像后进行 5.3.1 节的识别处理，当所有线程都处理完成后拼接处理后照片进行输出。并重新进入步骤 2。

本文使用处理终端详细配置如表所示。

表 5-3 处理终端配置

Table 5-3. Handle terminal

参数名		参数值
处理器	型号	i7 6700
	频率	3.4GHz
GPU	型号	GTX960
	容量	2GB
内存	型号	DDR4 2133
	容量	8GB
操作系统	型号	Window10
深度学习框架	类型	Pytorch0.4.0
处理语言	类型	Python3.6.2

5.4 本章小结

本章节简单描述了监控视角下人物识别系统的搭建，包括室内场景参数的选型，硬件设施如交换机和摄像头参数选择。其次基于摄像头摆布位置和数量，本文对人物识别流程进行详细规划，包含人脸识别，行人重识别以及基于时空关系的上下帧识别三种识别方式，提升系统识别准确率。其次基于处理终端容量大小和实时性需求，本文对多线程实时识别流程也进行规划，保证多线程工作时不会因为线程冲突问题导致识别无法进行。

第六章 总结与展望

6.1 总结

监控视角下人物识别是智能安防领域中的核心内容之一，在公安刑侦以及智慧城市方面发挥重要作用。由于硬件处理能力上升，大数据领域内的科技发展以及城市道路系统愈加复杂的影响，传统人物识别算法处理目前情况存在困难，包括如何从日益复杂的拍摄背景中准确分离行人目标，以及如何在监控视角下进行行人目标的准确识别。传统人物识别算法存在抗干扰能力差，无法对于大量行人目标无法有效区分问题，会降低公安刑侦效率以及智能安防的效果。

随着深度卷积网络的提出以及大量研究人员在该领域内的研究，深度学习广泛应用于计算机视觉的各个领域，在目标检测领域上 YOLO 系列目标检测算法大量应用于工业领域中国，在人物识别方面，基于深度卷积网络强大特征提取能力构建的识别模型展现出强大能力。本文利用深度卷积网络，对监控视角下人物识别算法存在问题进行研究，做出相关改进，弥补人物识别算法的行人目标分离以及行人目标准确识别的缺陷。本文主要工作如下：

（1）基于表征学习的深度卷积网络模型经过行人重识别数据集进行训练后，具备在监控视角下行人目标准确识别的能力，但是网络模型会存在对数据集的过拟合问题，无法进行有效泛化。基于表征学习的行人重识别网络模型可能存在的模型过拟合问题，本文提出对卷积网络模型进行中层特征扩展，将中层特征单独抽出，经过池化拼接等处理，与高层抽象特征一起输出，增加模型表征能力。该方法在可见光模态数据集上进行实验，实验结果表明中层特征扩展方法确实存在提升。

（2）为了增加监控视角下人物识别方法在光线不充足场景下识别能力，本文对可见光-热成像模态下行人重识别问题进行研究，针对跨模态行人重识别问题中多模态的情况，使用双流结构分别提取可将光模态和热成像模态行人图片特征，并且对传统可见光模态下困难三元组损失进行改进，扩展为跨模态双流困难三元组损失。其次在对损失函数改进的情况下，使用焦点损失增加困难样本在网络模型学习中的权重，提升网络模型训练效果。经过公开跨模态数据集 RegDB 进行实验，实验结果表明改进方法的有效性。

（3）由于深度学习方法在监控视角下人物识别算法的准确率上有提升，本文使用深度学习方法建立监控视角下人物识别系统。首先通过在室内环境下构建出摄像头网络模拟监控视角环境。其次采用 YOLOv3 目标检测网络对摄像头网络拍摄的照片进行行人目标检测并进行切割出行人目标，再采用行人重识别卷积网络进行人物识别。最后基于深度学习方法设计系统的识别流程图以及多线程识别流程图，搭建一套监控视角下人物识别系统并制作出一个可见光模态行人重识别数据集提供验证。本文的搭建的监控视角下人物识别方法为智能安防领域提供了一套可行的技术方案，并具有可拓展性。

6.2 展望

监控视角下人物识别算法会随着交通系统日益复杂和智能安防的产业升级经受更大的考验。受限于实验条件、隐私问题等多种问题，本文对行人重识别算法的研究以及监控视角下人物系统还存在不足以及提高的地方：

（1）本文在可见光模态下提出中层特征扩展方法中使用公开数据集和自建数据集共同进行验证，但是自建数据集相比公开数据集中还是缺乏公正性。自建数据集目前仅有 7610 个有效目标，相比传统公开可见光数据集中还存在不小差距，需要进一步收集可靠数据，增加自建数据集的公信力。

（2）本文在可见光-热成像模态中提出使用双流结构分别提取多模态行人图片信息，但是由于双流结构占用资源过多，需要对两个卷积网络模型进行权重更新，所以训练效率较慢，需要进行改进。

（3）本文在监控视角下人物识别系统中使用室内环境进行模拟搭建，但是真正的人物识别系统的场景不单单是室内小范围场景，更重要的是目前刑侦中常用的城市监控摄像头场景，大型活动中心等大型场景。本文提出的人物识别系统尚未在大场景下进行实验，所以无法预知在大型场景下的监控系统的效果。其次在实时性方面上，系统使用的多线程识别仅能刚好满足需求，需要对识别流程继续优化，提高效率。

参考文献

[1] 李云红, 魏妮娜, 张晓丹. 基于多方向 Gabor 滤波器的图像分割[J]. 国外电子测量技术, 2017, 36(3): 20-23.

[2] 赵泉华, 高郡, 李玉. 基于区域划分的多特征纹理图像分割[J]. 仪器仪表学报, 2015, 36(11): 2519-2530.

[3] Chen D, Cao X, Wen F, et al. Blessing of dimensionality: High-dimensional feature and its efficient compression for face verification[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. Piscataway, NJ: IEEE, 2013: 3025-3032.

[4] Wang X. Intelligent multi-camera video surveillance: A review[J]. Pattern recognition letters, 2013, 34(1): 3-19.

[5] Gheissari N, Sebastian T B, Hartley R. Person reidentification using spatiotemporal appearance[C]//2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06). Piscataway, NJ: IEEE, 2006, 2: 1528-1535.

[6] Bazzani L, Cristani M, Perina A, et al. Multiple-shot person re-identification by hpe signature[C]//2010 20th International Conference on Pattern Recognition. Piscataway, NJ: IEEE, 2010: 1413-1416.

[7] Lowe D G. Distinctive image features from scale-invariant keypoints[J]. International journal of computer vision, 2004, 60(2): 91-110.

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

[9] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[C]//Advances in neural information processing systems. Cambridge, MA: MIT Press, 2012: 1097-1105.

[10] Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, real-time object detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 779-788.

[11] Redmon J, Farhadi A. YOLO9000: better, faster, stronger[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. Piscataway, NJ: IEEE, 2017: 7263-7271.

[12] Redmon J, Farhadi A. Yolov3: An incremental improvement[J]. arXiv preprint arXiv:1804.02767, 2018.

[13] Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. Piscataway, NJ: IEEE, 2014: 580-587.

[14] Ren S, He K, Girshick R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks[C]//Advances in neural information processing systems. Cambridge, MA: MIT Press, 2015: 91-99.

[15] Jüngling K, Bodensteiner C, Arens M. Person re-identification in multi-camera networks[C]//CVPR 2011 WORKSHOPS. Piscataway, NJ: IEEE, 2011: 55-61.

[16] Layne R, Hospedales T M, Gong S, et al. Person re-identification by attributes[C]//Bmvc. Berlin: Springer, 2012, 2(3): 8.

[17] Karanam S, Li Y, Radke R J. Person re-identification with discriminatively trained viewpoint invariant dictionaries[C]//Proceedings of the IEEE international conference on computer vision. Piscataway, NJ: IEEE, 2015: 4516-4524.
[18] Li W, Zhu X, Gong S. Harmonious attention network for person re-identification[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. Piscataway, NJ: IEEE, 2018: 2285-2294.
[19] Sun Y, Zheng L, Yang Y, et al. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline)[C]//Proceedings of the European Conference on Computer Vision (ECCV). Berlin: Springer, 2018: 480-496.
[20] Wu A, Zheng W S, Lai J H. Robust depth-based person re-identification[J]. IEEE Transactions on Image Processing, 2017, 26(6): 2588-2603.
[21] Ye M, Wang Z, Lan X, et al. Visible Thermal Person Re-Identification via Dual-Constrained Top-Ranking[C]//IJCAI. Menlo Park, CA: AAAI, 2018, 1: 2.
[22] 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. Piscataway, NJ: IEEE, 2017: 5380-5389.
[23] Nguyen D T, Hong H G, Kim K W, et al. Person recognition system based on a combination of body images from visible light and thermal cameras[J]. Sensors, 2017, 17(3): 605.
[24] Fukushima K. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position[J]. Biological cybernetics, 1980, 36(4): 193-202.
[25] 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.
[26] Lin T Y, Goyal P, Girshick R, et al. Focal loss for dense object detection[C]//Proceedings of the IEEE international conference on computer vision. Piscataway, NJ: IEEE, 2017: 2980-2988.
[27] He K, Gkioxari G, Dollár P, et al. Mask r-cnn[C]//Proceedings of the IEEE international conference on computer vision. Piscataway, NJ: IEEE, 2017: 2961-2969.
[28] Liu W, Anguelov D, Erhan D, et al. Ssd: Single shot multibox detector[C]//European conference on computer vision. Berlin: Springer, 2016: 21-37.
[29] Sermanet P, Eigen D, Zhang X, et al. Overfeat: Integrated recognition, localization and detection using convolutional networks[J]. arXiv preprint arXiv:1312.6229, 2013.
[30] Oreifej O, Mehran R, Shah M. Human identity recognition in aerial images[C]//2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2010: 709-716.
[31] Ma B, Su Y, Jurie F. Local descriptors encoded by fisher vectors for person re-identification[C]//European Conference on Computer Vision. Berlin: Springer, 2012: 413-422.
[32] Matsukawa T, Okabe T, Suzuki E, et al. Hierarchical gaussian descriptor for person re-identification[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. Piscataway, NJ: IEEE, 2016: 1363-1372.
[33] Schroff F, Kalenichenko D, Philbin J. Facenet: A unified embedding for face recognition and clustering[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. Piscataway, NJ: IEEE, 2015: 815-823.
[34] Varior R R, Haloi M, Wang G. Gated siamese convolutional neural network architecture for human re-identification[C]//European conference on computer vision. Berlin: Springer, 2016: 791-808.

[35] Chen W, Chen X, Zhang J, et al. Beyond triplet loss: a deep quadruplet network for person re-identification[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2017: 403-412.

[36] Cheng D, Gong Y, Zhou S, et al. Person re-identification by multi-channel parts-based cnn with improved triplet loss function[C]//Proceedings of the iEEE conference on computer vision and pattern recognition. Piscataway, NJ: IEEE, 2016: 1335-1344.

[37] Hermans A, Beyer L, Leibe B. In defense of the triplet loss for person re-identification[J]. arXiv preprint arXiv:1703.07737, 2017.

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

[39] Yu Q, Chang X, Song Y Z, et al. The devil is in the middle: Exploiting mid-level representations for cross-domain instance matching[J]. arXiv preprint arXiv:1711.08106, 2017.

[40] Zheng L, Shen L, Tian L, et al. Scalable person re-identification: A benchmark[C]//Proceedings of the IEEE international conference on computer vision. Piscataway, NJ: IEEE, 2015: 1116-1124.

[41] Chang X, Hospedales T M, Xiang T. Multi-level factorisation net for person re-identification[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 2109-2118.

[42] Hu J, Shen L, Sun G. Squeeze-and-excitation networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. Piscataway, NJ: IEEE, 2018: 7132-7141.

[43] 宋婉茹, 赵晴晴, 陈昌红, 等. 行人重识别研究综述[J]. 智能系统学报, 2017, 12(6): 770-780.

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

[45] Matsukawa T, Suzuki E. Person re-identification using cnn features learned from combination of attributes[C]//2016 23rd international conference on pattern recognition (ICPR). Piscataway, NJ: IEEE, 2016: 2428-2433.

[46] Ye M, Lan X, Li J, et al. Hierarchical discriminative learning for visible thermal person re-identification[C]//Thirty-Second AAAI conference on artificial intelligence. Menlo Park, CA: AAAI, 2018.

[47] Nguyen D T, Hong H G, Kim K W, et al. Person recognition system based on a combination of body images from visible light and thermal cameras[J]. Sensors, 2017, 17(3): 605.

[48] Zhu Y, Yang Z, Wang L, et al. Hetero-Center Loss for Cross-Modality Person Re-Identification[J]. Neurocomputing, 2019.

[49] Lin L, Wang G, Zuo W, et al. Cross-domain visual matching via generalized similarity measure and feature learning[J]. IEEE transactions on pattern analysis and machine intelligence, 2016, 39(6): 1089-1102.

[50] Liao S, Li S Z. Efficient psd constrained asymmetric metric learning for person re-identification[C]//Proceedings of the IEEE International Conference on Computer Vision. Piscataway, NJ: IEEE, 2015: 3685-3693.

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

致　谢

研究生的时光愉快而充实，弹指之间我已经在研究生这条路上拼搏了三年，三年的时光对我的影响很深，在这人生最美好的三年里，我从一开始对专业领域内的懵懂无知，到现在能够在专业前沿思考求索，寻找新的思路，在这进步的过程中受到了导师的勤奋指导，受到了同学师兄弟间的相互帮助，还受到了父母亲人们的鼓励。研究生的三年间，我曾经对前路的迷茫而沮丧，曾经对研究道路上的困难而无能为力，但是我也享受过收获科研成果的喜悦，享受过被人认同的充实感。这三年来我受到了太多人的帮助，在论文完成之际，我在这里向那些帮助，支持，鼓励我的人送上最诚挚的谢意！

首先，非常感谢我的导师赵云波老师，赵老师为我们科研生活提供了非常好的条件，带领我们走向研究生活，并且让我们自由发挥自己的想法。当我在科研道路上遇到困难时，赵老师一直会耐心的帮忙和引导我，指出需要改进和前进的方向，让我更好的解决困难。从赵老师身上我学到了如何理性去分析科研问题，如何从理论和当前局面推导出问题关键并更好解决问题。赵老师认真工作，努力创新的科研态度让我感触很深，希望在今后的人生发展的道路上我会一直保持着这种态度，努力提升自己。

其次，感谢实验室的师兄弟们，在科研上以及生活上师兄们对我帮助非常大，常常会帮助我解决问题，提供技术上以及生活上的指导，让我受益匪浅，帮助我更好的研究问题。其中，特别感谢带我进入深度学习领域的林建武同学，我们一起参与赵老师的横向课题项目，一起研究深度学习图像领域的科研问题，一起参加之江实验室的合作项目，一起规划毕业后未来的生活，这些共同奋斗的回忆会让我铭记于心。

还要感谢我的家人们，他们一直是我坚强的后盾，让我在科研路上不懈努力，坚持向前。他们的关心以及支持是我三年来最感动的回忆，谢谢他们为我付出的一切！

最后，衷心感谢各位专家，学者和老师抽出宝贵的时间对本文的评阅！

作者简介

1 作者简历

1995 年 6 月出生于广东省惠州市。

2013 年 9 月——2017 年 6 月，浙江科技学院学院自动化及电气工程学院电气工程及其自动化专业学习，获得工学学士学位。

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

2 攻读硕士学位期间发表的学术论文

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

3 参与的科研项目及获奖情况

[1] 基于资源调度和预测控制的无线网络化控制系统的联合设计. 中国国家自然科学基金项目(61673350).

4 发明专利

[1] 赵云波, 李灏, 林建武. 一种基于中层特征扩展卷积网络的农作物病害分析方法. 中国 2019 1 0950143.1 [P]. 2019-10-08.

[2] 赵云波, 林建武, 李灏. 一种基于行人重识别的员工特定行为记录方法. 中国 2019 10178684.7 [P]. 2019-03-11.

[3] 赵云波, 林建武, 李灏, 宣琦. 一种基于密集网络的多任务卷积神经网络顾客行为分析方法. 中国, 2018 1 1317143.X [P]. 2018-11-07.

[4] 赵云波, 李灏, 林建武. 一种基于多线程的多摄像头实时检测方法. 中国, 2018 1 1197765.3 [P]. 2018-10-15.

[5] 赵云波, 李灏, 林建武, 宣琦. 一种基于轻量化多任务卷积神经网络的导购行为分析方法. 中国, 2018 1 1197730.X [P]. 2018-10-15.

[6] 赵云波, 林建武, 李灏. 一种基于人脸识别与行人重识别的特定目标跟踪方法. 中国, 2018 1 1196063.3 [P]. 2018-10-15.

[7] 赵云波, 林建武, 李灏, 宣琦. 一种基于 yolo 和多任务卷积神经网络的导购消极行为监控方法. 中国, 2018 1 1197781.2 [P]. 2018-10-15.

学位论文数据集

密　级*	中图分类号*	UDC*	论文资助
公开	TP391	004.8	
学位授予单位名称	学位授予单位代码	学位类型*	学位级别*
浙江工业大学	10337	工学硕士	全日制学术型硕士
论文题名*	监控视角下基于深度学习的人物识别方法研究		
关键词*	深度学习，行人重识别，表征学习，跨模态，困难三元组损失		论文语种*
并列题名*	Deep Learning Based Person Recognition Methods under Surveillance Cameras		中文
作者姓名*	李灏	学　号*	2111703056
培养单位名称*	培养单位代码*	培养单位地址	邮政编码
浙江工业大学 信息工程学院	10337	杭州市潮王路 18 号	310014
学科专业*	研究方向*	学　制*	学位授予年*
控制科学与工程	计算机视觉	3 年	2020 年
论文提交日期*	2020 年 6 月		
导师姓名*	赵云波	职　称*	教授
评阅人	答辩委员会主席*	答辩委员会成员	
盲审	石崇源	宣琦，赵云波，陈晋音，黄亮	
电子版论文提交格式：文本（ ）图像（ ）视频（ ）音频（ ）多媒体（ ）其他（ ）			
电子版论文出版（发布）者	电子版论文出版（发布）地		版权声明
论文总页数*	52 页		
注：共 33 项，其中带*为必填数据，为 22 项。			