浙江工业大学

本科毕业设计论文

题目：基于深度学习的人物识别及实现

作者姓名 高凯

指导教师 赵云波 教授

专业班级 自动化 1501 班

学　　院 信息工程学院

提交日期 2019 年 6 月 10 日

浙江工业大学本科毕业设计论文

基于深度学习的人物识别及实现

作者姓名：高凯

指导教师：赵云波教授

浙江工业大学信息工程学院

2019 年 6 月

Dissertation Submitted to Zhejiang University of Technology for the Degree of Bachelor

Research on Person Recognition and Implementation Based on Deep Learning

Student: Kai Gao

Advisor: Professor Yunbo Zhao

College of Information Engineering
Zhejiang University of Technology

June 2019

浙江工业大学

本科生毕业设计(论文、创作)诚信承诺书

本人慎重承诺和声明：

1. 本人在毕业设计（论文、创作）撰写过程中，严格遵守学校有关规定，恪守学术规范，所呈交的毕业设计（论文、创作）是在指导教师指导下独立完成的；

2. 毕业设计（论文、创作）中无抄袭、剽窃或不正当引用他人学术观点、思想和学术成果，无虚构、篡改试验结果、统计资料、伪造数据和运算程序等情况；

3. 若有违反学术纪律的行为，本人愿意承担一切责任，并接受学校按有关规定给予的处理。

学生（签名）：高凯

2019 年 6 月 1 日

浙 江 工 业 大 学

本科生毕业设计（论文、创作）任务书

专 业 自动化 班 级 自动化 1501 学生姓名/学号 高凯/201503080108

一、设计（论文、创作）题目：

基于深度学习的人物识别及实现

二、主要任务与目标：

1. 阅读相关文献，了解本领域研究现状；
2. 归纳并选择合适的方法进行人物识别；
3. 进行人物识别；
4. 撰写毕业论文.

三、主要内容与基本要求：

由于安防系统以及人工智能的快速发展，会遇到需要快速识别人物的场景，例如：警察 通过视频回溯系统寻找犯人，公共场所对丢失儿童的识别等等。这类场景中传统的方法存在 效率低，准确度差的情况。因此对此课题研究有较大社会意义。

本课题旨在研究人工智能在人物识别领域中的应用，利用目前火热的深度学习神经网络 技术进行识别研究。

四、计划进度：

2019 年开学前，收集相关资料文献，学习相关知识，完成外文翻译、文献综述；熟悉课题，做好开题准备 第 1-3 周，完成开题报告，参加开题交流 第 4-8 周， 初步开发识别算法，接受中期检查 第 9-14 周， 完善识别算法并实现，撰写毕业论文 第 15 周， 修改毕业论文，参加毕业答辩，提交相关文档资料

五、主要参考文献：

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” pp. 1097–1105, 2012. [2] S. Ren, K. He, R. Girshick, and J. Sun, “Faster r-cnn: Towards real-time object detection with region proposal networks,” pp. 91–99, 2015. [3] F. Schroff, D. Kalenichenko, and J. Philbin, “Facenet: A unified embed-ding for face recognition and clustering,” pp. 815–823, 2015.

任务书下发日期 2018 年 12 月 25 日

设计（论文、创作）工作自 2018 年 12 月 25 日 至 2019 年 6 月 4 日

设计（论文、创作）指导教师 赵云波

系主任（专业负责人）

主管院长

基于深度学习的人物识别及实现

摘　要

随着监控摄像头技术不断发展，在小区、广场、各大街道等众多公共场所都安装了摄像头。依靠人工监控这些摄像头需要耗费大量人力物力，如何自动检测并识别监控视角下的人物成为了当今研究的热点之一。随着人工智能的不断发展，基于深度学习的人物识别方法受到了研究者们的青睐。

本文的主要研究内容是基于深度学习的人物识别及实现，是应对警务和安防的需求而提出的，用于在视频流中捕获特定的人物，如走失的老人儿童、可疑的犯罪分子等。在实际监控视角下，存在无法拍到行人正脸的情况，这就导致无法使用人脸识别方法对行人进行识别。因此，本文采用行人重识别的方法进行代替。实验首先对采集得到的视频进行行人检测，将检测到的行人放入查询图库中，然后分别基于搜索目标和查询图库进行特征提取，通过近似度匹配得到相似的人物。本文就行人特征的提取设计了一种兼顾全局特征和局部特征的多粒度网络，结合行人检测与行人重识别实现了基于深度学习的人物识别。

本论文的主要工作如下：

1. 综述了基于深度学习的人物识别算法的研究现状和目标检测常用算法，了解这些算法的实际适用环境。

2. 提出了一种兼顾全局特征和局部特征的多粒度网络用于特征提取，在 Market1501 数据集上测试实现了 95.22%的 Rank-1 和 93.23%的 mAP。

3. 结合行人检测与行人重识别，实现了监控视角下对某个特定行人的识别。

关键词：深度学习，人物识别，目标检测，行人重识别，多粒度网络

RESEARCH ON PERSON RECOGNITION AND IMPLEMENTATION BASED ON DEEP LEARNING

ABSTRACT

With the development of surveillance camera technology, cameras have been installed in many public places such as communities, squares, and major streets. Relying on manual monitoring of these cameras requires a lot of manpower and material resources. How to automatically detect and identify people in the monitoring perspective has become one of the hotspots of today's research. Due to the development of artificial intelligence, the method of character recognition based on deep learning has been favored by researchers.

The main research content of this paper is based on the deep recognition of character recognition and implementation, which is proposed to meet the needs of police and security. It’s used to capture specific characters in the video stream, such as lost elderly children and suspicious criminals. In the actual monitoring perspective, there is a situation in which a pedestrian's face cannot be photographed, which makes it impossible to identify a pedestrian using a face recognition method. Therefore, this paper uses the method of pedestrian recognition to replace it. The experiment firstly performs pedestrian detection on the captured video, puts the detected pedestrians into the query library, and then extracts features based on the search target and the query library respectively, and obtains similar characters through approximation matching. In this paper, a multi-granularity network with global and local features is designed for the extraction of pedestrian features. Combining pedestrian detection and pedestrian recognition, the character recognition based on deep learning is realized.

The main work of this paper is as follows:

1. The current research status of target recognition algorithms based on deep learning and common algorithms for target detection are reviewed, and the practical application environment of these algorithms is understood.

2. A multi-granularity network with both global and local features is proposed for feature extraction. 95.22% of Rank-1 and 93.23% of mAP are trained on the Market1501 dataset.

3. Combined with pedestrian detection and person re-identification, this paper realizes the recognition of a specific pedestrian from the perspective of monitoring.

Key words：deep learning, character recognition, object detection, pedestrian re-identification, multi-granular network

目　录

第1章 绪 论

1.1 课题研究背景及意义

近年来，随着监控建设的飞速发展，摄像头的安装已经遍布在各个小区、街道、广场等公共场所，以充分发挥视频图像信息效能为核心的视频警务模式已悄然兴起。这些摄像头源源不断地采集视频和图像信息，作为警务工作的探查基础。

通过摄像监控对特定人物进行识别、定位、查找的需求逐年增加。数据显示，儿童走失案件多发于公园、游乐场、车站等公共场所，这些场所人员走动密集，容易发生拐卖事件。老年人由于记忆衰退、表达能力欠缺、行动不便等因素，走失时间过长容易造成死亡。通过监控迅速找到这些走失人员具有极大的社会效益。另一方面，对于犯罪分子的抓捕也需依靠城市街道的视频监控。一些犯罪分子善于伪装、躲避摄像监控，经常通过佩戴帽子、墨镜、口罩等物品对人脸特征进行伪装。如何通过监控视频快速发现这些可疑人员并追踪其行动轨迹，是刑侦工作的难点。

传统的视频监控解决了视频的存储和回放，以及视频流的互联互通，但依旧无法准确识别、定位和查找目标信息，往往需要依靠大量的警务人员时刻紧盯监控或者回放所有相关视频录像进行线索的查找，这耗费了大量的人力和时间，容易给目标人物进行地点转移制造机会。另外，人力进行监控也会因为疲劳、疏忽而遗漏一些稍纵即逝的重要信息，这会对走失人员的寻回以及犯罪分子的抓捕工作增加难度。

随着人工智能技术的兴起，人物识别技术得到了巨大的发展。依靠领先的人工智能算法在一线警务实战应用中的深度结合，已经开发出较为成熟的人脸识别系统。人脸识别是通过摄像头采集包含人脸的图像或者视频，并能够自动检测和跟踪人脸并识别人物信息，是一种基于人的脸部特征信息进行身份识别的生物识别技术，是构建立体化现代化社会治安防控体系的重要支撑之一。但在实际监控视角无约束环境下，因为摄像头分辨率不高或者摄像头角度有差异的缘故，一般不能直接获得较清楚的面部细节，尤其对于犯罪分子的面部伪装情况，此时进行

人脸识别会有较大的误判率。研究一种可以代替人脸识别的人物识别技术成为此类刑侦工作的突破点。

1.2 基于深度学习的人物识别研究现状综述

深度学习是 Hinton 等学者在 2006 年提出[1]，最初是源于人工神经网络的研究。深度学习是一种可以实现层级模式提取和识别的方法论，它通过组合低层特征形成更加抽象的高层特征以发现数据的特征表示规律。深度学习，特别是卷积神经网络(Convolutional Neural Networks, CNNs)近年来在图像识别领域取得了巨大突破，这表明深度学习在特征提取上存在较大优势。

然而，基于深度学习的人物识别方法在监控视频中仍有许多问题。一方面，基于深度学习的人物检测系统存在鲁棒性差，资源占用多等问题[2]。另一方面，实际监控视角下的背景十分复杂，光线强度也会随着时间发生改变，行人之间容易互相遮挡或被其他物体遮挡，而且行人可能存在服装统一等问题。另外，监控视角的拍摄角度不同，图像分辨率的要求也不同，这些都大大限制人物识别技术的发展，严重影响了人物识别算法的效果。同时，监控视角下的人物图片会存在变形、模糊、无法提取特征信息等问题。低质量的照片会直接影响模型的精度，从而导致整个系统性能的下降。如何综合利用深度学习的提取特征优势解决监控视频里人物图像的不确定性是研究的难点,也是人物识别工作中需要迫切解决的问题。

在实际应用中，由于视频中人物移动的复杂性、随机性，以及监控视角下的人物形变等问题，如何提高在监控视角下人物识别的准确度是本文研究的问题。目前的研究重点主要集中于人脸识别和行人重识别两个方面。

1.2.1 基于深度学习的人脸识别研究

人脸识别(Face Recognition)，即通过摄像头采集包含人脸的图像或者视频，并能够自动检测和跟踪人脸并识别人物信息，是一种基于人的脸部特征信息进行身份识别的生物识别技术，目前已广泛应用在各个领域。通常地，人脸识别包括人脸检测、图像预处理（包括人脸对准、人脸矫正等）和人脸识别（确认人物身份）

三个步骤。

人脸检测和对齐对于许多人脸应用来说至关重要，然而由于各种原因，比如遭受遮挡、光线较弱或者其他问题，在无约束环境下的人脸检测和对准将会有相当大的难度。最近研究表明，深度学习在解决脸部检测和对齐方面有着巨大贡献。特别地，卷积神经网络(CNNs)在人脸识别方面已经取得了巨大进展[5-8]。在DeepFace[4]和 DeepID[8]中，人脸识别被视为多分类的问题，并且第一次引入深度卷积网络来学习具有多重类标的数据集的特征。文献[4]提出了一个 DeepFace 系统，使用 3D 建模技术使人脸重新对准，经过一个 9 层的 CNN 获得了人脸的表达。该方法也在 LFW(Labeled Face in Wild，户外标记人脸数据集)上得到了 97.35%的人脸验证精度。DeepID2[8]采用识别和验证信号来实现更好的特征嵌入。最近的工作 DeepID2 +[5]和 DeepID3[6]进一步探索了先进的网络结构，以提高识别性能。

其他算法也证实了深度 CNN 对人脸识别的有效性。文献[17]用深度卷积神经网络进行人脸局部属性检测，然后综合各个脸部部件（头发、眼睛、鼻子、嘴和胡子）进行打分机制最终得到人脸检测结果。虽然其打分机制对于人脸部分遮挡有较好的解决效果，但在实际检测过程中，因为网络结构较为复杂可能需要耗费大量的时间。文献[18]提出了一种多分辨率的级联卷积神经网络用于高速情况下人脸的检测，其中额外增加了一定的计算花费用于检测更小的人脸，但可以明显地提高识别率，且在整个过程中该开销可忽略不计。文献[19]使用 Multi-Task Learning (MTL)进行人脸核心位置点的检测，提出了 Tasks-Constrained Deep Convolutional Network (简称 TCDCN)。该文献在进行人脸特征点的检测时通过添加辅助的信息进行更好的人脸特征点定位，包括性别，是否戴眼镜等，但是这种方法对于弱面部检测具有一定的限制。文献[3]提出了一个深度级联的多任务框架，该框架在检测和对准之间建立联系，通过 3 个精巧的 CNN，以粗略到精细的方式估计人脸和地标位置，通过一个新的在线硬样本挖掘策略提高实践性能，在面部对齐的 AFLW 基准测试中实现了卓越的精度。

另一方面，为了训练出更具内聚性的特征，研究者开始对损失函数进行研究，旨在学习到的特征具有更好的泛化能力和辨别能力。通常，在深度卷积网络中，

多层感知器网络连接着 softmax 损失[4,21]，然而，最近的研究发现传统的 softmax 损失不足以获得分类的辨别力[22,23,8-10]。为了激发更好的辨别性能，文献[7]和[10]分别提出了 constrastive loss 和 triplet loss 损失函数。文献[9]在原有 softmax loss 适用的情况下添加了 center loss 损失函数，网络收敛速度大大得到了提高，且不需要构造大量的训练对。文献[22]通过为每个类添加角度约束来提出大边缘 softmax（L-Softmax）以改进特征判别。文献[23]利用角度 softmax（A-Softmax）通过归一化权重来改善 L-Softmax，从而在一系列开放式人脸识别基准上实现了更好的性能[24,25]。文献[20]提出了一种新的面向深度人脸识别的大边缘余弦损失函数(LMCL)，该函数通过对特征和权向量进行 L_2 归一化以消除径向变化来重新设计 softmax loss 作为余弦损失，在此基础上引入余弦边缘以进一步最大化角度空间中的决策边界。于是，通过归一化的优势和余弦决策边缘的最大化，实现了最小的类内方差和最大的类间方差。

另外，为了解决侧脸识别效果不佳的问题，港中大和商汤科技等[11]研究出了一种深度表示空间中在正脸和侧脸间通过等价映射建立函数关系的方法。这种方法的所需的计算量较小，但侧脸的识别效果得到了较大的提升。

1.2.2 基于深度学习的行人重识别研究

行人重识别(Person Re-identification)简称 ReID，也称为行人再识别。在监控视角下因为摄像头分辨率不高或者摄像头角度有差异的缘故，一般不能直接获得较清楚的面部细节，此时进行人脸识别会有较大的误判率，行人重识别技术就起到了一个很重要的替代作用。行人重识别(ReID)，在计算机视觉领域被形象地认为是针对监控视频的检索问题，关键是判定给定视频中的特定行人有无出现在其他视频中。目前该问题有以下几种解决方案：

① 基于表征学习的 ReID 方法

得益于 CNN 的快速发展，基于表征学习(Representation Learning)的行人重识别是一类比较常用的方法。研究者把该问题看做分类问题或者验证问题。分类问题即利用行人 ID 或属性作为标签来训练模型，而验证问题即给定两张人物图片让 CNN 自动学习判断是否属于同一个人。文献[12]利用 Classification/

Identification loss 和 verification loss 来训练网络，但是也有论文[13]认为仅凭行人的 ID 信息不能够学习出一个十分鲁棒的模型。它们额外添加了行人的其他特征，如性别、服装、姿态等属性，来增强模型对新鲜样本的适应能力，实践验证这种方法是可靠的。目前，基于表征学习的行人重识别方法因其较好的鲁棒性成为解决该类任务的一个基准。

② 基于度量学习的 ReID 方法

度量学习(Metric Learning)广泛应用于图像检索领域。该方法的基本思路是最小化类内方差和最大化间方差，常常通过网络的损失函数来实现该基本思路。度量学习中常用的损失函数有三元组损失(Triplet loss)[13,14]、四元组损失(Quadruplet loss)[26]、对比损失(Contractive loss)[14]、边界挖掘损失(Margin sample mining loss, MSML)[28]、难样本采样三元组损失(Triplet hard loss with batch hard mining, TriHard loss)[27]。

③ 基于局部特征的 ReID 方法

由于在全局特征上遇到了瓶颈，研究者们开始在局部特征上寻找突破口。基于局部特征的 ReID 方法解决关键是通过分割图像、定位骨架关键点以及对准姿态等方法进行识别。图像垂直切割[29]是一种常用的方式用来提取局部特征，但其对于图像对齐的要求较高。为了解决这种图像未对齐的情况，文献[30]首先通过姿态估计的方法预测出行人的 14 个关键点，再通过仿射映射把相符的关键点进行对准。文献[16]也利用了相同数量的人体关键点，并直接通过这些关键点找出感兴趣区域(Region of Interest，ROI)，将这些 ROI 和原始图片通过 CNN 进行特征提取。文献[31]提出了一种全局和局部对准进行特征的描述算子(Global-Local-Alignment Descriptor, GLAD)，用于处理姿态不同的问题。

④ 基于视频序列的 ReID 方法

由于从单帧图像中提取信息有限，研究者将目光投向视频序列相邻帧之间的信息。基于视频序列的 ReID 方法基本思路是通过卷积神经网络进行空间特征提取并通过递归循环网络(Recurrent neural networks, RNN)来进行时序特征的提取，最终利用这些特征来训练网络。累积运动背景网络(Accumulative motion context

network, AMOC) 是处理该问题较为典型的方法[32]。使用 AMOC 的关键是网络要同时获取序列图像特征和光流运动特征。该网络采用了分类损失和对比损失来训练模型，在添加了包含多帧信息的图像运动特征后 ReID 的准确度得到了提高。另一篇文献[33]则明确提出单帧图像在受到遮挡的时候可用多帧信息来补充信息。

⑤ 基于 GAN 造图的 ReID 方法

行人重识别有一个很大的问题就是数据获取困难。目前最大的 ReID 数据库也不过几千个 ID，远远达不到拓展研究的程度。文献[34]首次提出利用 GAN 来解决 ReID 数据少的问题。这篇文章提出了一个标签平滑的方法，将随机生成的图像加入训练中，但生成的图像质量不够高，已经无法达到现在希望达到的效果。文献[35]则提出能控地进行 GAN 造图，对文献[34]中的标签平滑做出了一定的改进。总的来说，GAN 造图解决了 ReID 数据获取困难的问题。

1.3 基于深度学习的目标检测研究综述

目标检测(Object Detection)，即确定某些图像中是否存在来自给定类别(例如人、汽车、狗等)的任何目标实例，如果存在，则通过 bounding box[36]返回每个对象在图像中的大小和坐标[37]。

目标检测是计算机视觉(Computer Vision)中最基本也是最具挑战性的问题之一：之所以基本，是因为目标检测是解决许多更复杂或更高级的计算机视觉问题的基础，如行人重识别等；之所以具有挑战，是因为每个类别内的差异、目标实例的差异、无约束的环境等因素都容易对目标检测结果造成干扰。目前，在人工智能信息技术等领域，目标检测已具有相当广泛的应用，包括机器人视觉、无人驾驶、人机交互、智能视频监控和增强现实、消费电子等。传统的目标检测更多得在于检测特定的类别(如人脸、行人)。随着深度学习技术的发展，学者们开始转向研究通用目标检测系统的开发。

2012 年，深度学习第一次应用于目标检测领域。Hinton 和 Alex Krizhevsky 等人设计了一个深度卷积神经网络——AlexNet [38]，在大规模视觉识别竞赛 ImageNet [39]中实现了 57.1%的精度和 80.2%的 top-5，这个指标在当时可谓十分

亮眼。AlexNet 的提出预示着深度学习在目标检测领域的正式登场。从那时起，许多基于深度学习的检测器不断开发出来。

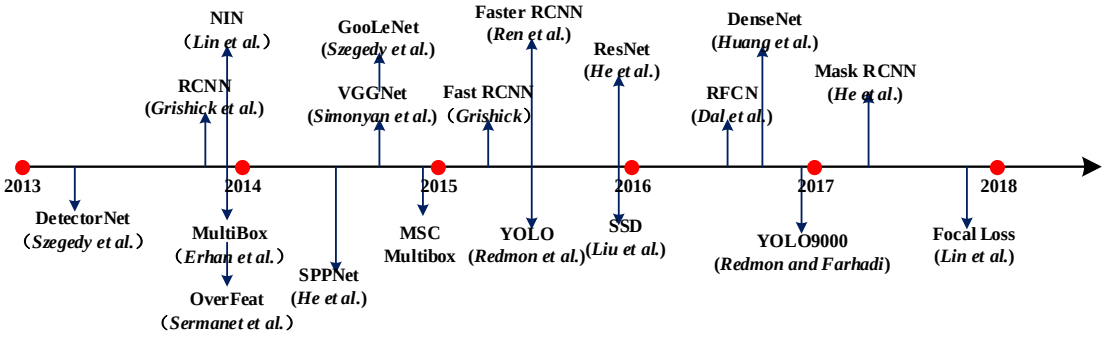

图 1-1　基于深度学习的目标检测框架

图 1-1 记录了自 2012 年深度学习应用于目标检测领域主要检测的提出时间。在一定程度上，这些检测器的提出要归功于 GPU 计算资源和大规模数据集[52](如 ImageNet、COCO[53]等)的可用性。

高效的检测框架，如级联、共享特性计算等，对于减小计算复杂度有着较为关键的作用。以上检测器在框架上大致可分为两种：①两级检测框架，基于这种框架的检测器首先针对图像中的目标检测物体预先提出候选区域；②单级检测框架，基于这种框架的检测器不预先提出候选区域，基于该框架的算法检测器速度普遍更快。针对这两种框架，下面分别归纳了相应的具有典型意义的检测器。

1.3.1　基于区域提名的目标检测算法

① R-CNN

R-CNN(Region Convolutional Neural Networks)，是最先通过深度学习进行目标检测的检测器之一。Girshick 等人首次通过探索 CNN 的通用目标检测性提出了 R-CNN[40]，其框架思路如下：

1. 使用 Select Search 技术生成类别独立的候选区域，即所谓的区域提名；
2. 将提名区域从图片中剪辑并变形成面积相同的区域，将其作为输入。在 ImageNet 大型数据集上进行预训练操作并微调在 VOC[50-51]数据集上训练

的权重；

3. 使用CNN对每个提名区域用一个4096维的特征向量表示，然后采用SGD训练一系列SVM分类器，替换softmax分类器进行类别预测；
4. 训练一个线性的边界框回归模型用于边界框的定位。

虽然 R-CNN 实现了较高质量的目标检测，但其训练步骤较为繁琐，训练、测试的过程也十分缓慢，加之 SVM 模型难以优化，训练十分占用存储空间。

② Fast R-CNN

Fast R-CNN 由 Girshick 提出[41]，其大致框架与 R-CNN 一致但在 R-CNN 上做出了不少的改进。首先，为了避免经过预训练的卷积神经网络进行检测的步骤，Fast R-CNN 提出了一个 RolPooling 层，是在最后一个卷积层和第一个全连接层之间添加的池化层，这一改进策略使得整个网络可以实现端到端的学习；同时，Fast R-CNN开发了一个流线型的训练过程，该过程使用Softmax分类器代替SVM，以 multi-task loss 的方式进行边界框回归。另外，Fast R-CNN 采用了截断的 SVD 分解方式来加速网络。由于 Fast R-CNN 实现了端到端的学习，因此训练与测试的速度比 R-CNN 提高了很多，训练速度提高了 3 倍，测试速度提高了 10 倍。

③ Faster R-CNN

虽然 Fast R-CNN 显著提高了训练和测试的速度，但仍然依赖于外部区域的提名。而研究显示，CNN在CONV层定位对象的超高能力会在FC层被减弱[42]，因此可以利用 CNN 替换选择性搜索方法产生区域提名。于是，Ren 等人建设性地提出了 Faster R-CNN 的框架[43]，通过在卷积神经网络后加入区域生成网络(Region Proposal Network，RPN)作为分支网络候选框的提取，取代之前时间开销较大的选择性搜索方法。于是，区域提名、分类、回归等操作可以一起共享卷积特征，进一步提升了速度。RPN 实际上是一种全卷积网络(FCN) [44]，用于提取候选框，使用的本质是滑动窗口，可以针对提名区域进行端到端的训练。在 Faster R-CNN 中，检测问题的基础步骤（候选区域的提名，特征的提取，分类，边框坐标的定位）被合并于一个框架之内。

图 1-2　R-CNN、Fast R-CNN、Faster R-CNN 的框架变化

1.3.2　基于端到端学习的目标检测算法

① YOLO 系列

YOLO 的名字来源于论文“ You Only Look Once”，是 Redmon 等人[45]提出的。不同于 R-CNN 系列，YOLO 最大的创新是放弃了对候选区域的提名，而是采用端到端学习的过程对图像进行直接处理。YOLO 将整个图像的全局特征作为输入，直接在输出层回归边界框的坐标和属于的类别。尤其是 YOLO 把图像平均划分成了数目为 S×S 的网格。这些网格每个都可以预测是否有对象属于 C 个类别中的一个，并在最后返回 B 个边界框位置和这些边界框的 confidence 分数。于是，整个图像最后可以用 S×S×(5B + C)的 tensor 来表示。由于 YOLO 在划分网格时相对“随意”，而且划分好的每个图像格子只能用来检测一个对象，所以当某个对象在图像中占画面较小时，容易导致无法被检测到。另外，某些物体特殊的长宽比导致了 YOLO 算法泛化能力较差。基于此，Redmon 和 Farhad 提出了改进版本的 YOLO 算法(YOLOv2[46])，YOLOv2 用 DarkNet19 网络替代了第一版本的 GooleNet[47]网络并添加了一些先进策略用于改进，如大批量正则化[48]、移除 FC 层、使用 kmeans 聚类和多任务训练学习得到的 anchor boxes 等。另外，Redmon 和 Farhad 也介绍了 YOLO9000。YOLO9000 可以通过一种联合优化方法实时监测超过 9000 种类别的目标。

② SSD

SSD 是 Single Shot Detector 的简称，由 Liu 等人[49]提出，在不以减小检测精度为代价的前提下保证检测的速度，因此它的训练和测试更快。类似于 YOLO 算法，SSD 按照不同的 scale 和 ratio 生成 k 个候选框并预测若干个 bounding box 和

这些 bounding box 中存在对象的分数，然后经过非极大值抑制确定最后可以用来表示检测结果的边界框。在 SSD 中，基础网络是全卷机的神经网络，在基础网络的末端加入若干个尺寸不断减小的卷积层用于辅助。另外，SSD 通过浅层来检测在图像中占比很小的物体，因为这些浅层分辨率更高。为了能够检测不同大小的目标对象，SSD 在卷积特征图谱的多个尺度上执行检测操作，每特征图谱都预测关于物体类别的 21 个置信度得分和 4 个框偏移量。

1.4 主要研究内容

本文的研究内容是针对当下警务需求提出一种基于深度学习的人物识别算法，用于实现在搜集视频中定位追踪所查询的特定人物。本文主要章节内容如下：

第一章：绪论。本章介绍了课题研究背景及意义，综述了深度学习在人物识别领域的研究现状，主要介绍了人脸识别和行人重识别两个方面的方法研究。基于本课题背景，行人重识别将作为本论文的研究重点。另外，对于行人重识别的预处理步骤目标检测，以框架的形式分类对主流的目标检测器进行综述。最后给出了全文的主要研究内容安排。

第二章：行人检测与行人重识别。本章首先给出了行人重识别算法用来训练测试的数据集和主要性能评价指标。然后，给出了主流的损失函数计算公式。对于行人检测，本章就基于实验用到的 YOLOv3 行人检测算法进行了较为详细的分析。最后，本章给出了行人重识别的实现思路。

第三章：多粒度网络的提出。本章首先引入了行人重识别问题采用的基础网络，以基础网络为启发，提出了一种兼顾全局特征和局部特征的多粒度网络。

第四章：基于多粒度网络的人物识别算法实现。本章首先给出了实验需要的基础库和算法实现流程，然后利用 yolov3 检测行人制作查询图集。对行人重识别算法分别测试基础网络和提出的多粒度网络，并给出比较。最后，分别给出静态图库下行人识别的 top10 结果和动态视频下的识别截图

第五章：总结与展望。本章简要总结了本文进行的工作并对现存的一些问题进行未来展望。

1.5 本章小结

本章作为绪论，简要介绍了本课题的研究背景意义，综述了现阶段基于深度学习的人物识别的研究现状以及基于深度学习的通用目标检测器的研究现状并给出全文的章节安排。

第 2 章　行人检测与行人重识别

2.1　数据集和评价指标

2.1.1　数据集

目前的行人重识别研究基本上是基于公开数据集的研究。这些数据集多为图片库的形式。通过在某一个特定的区域内采集几个摄像头的行人图像，然后进行人工标注或自动标注。这些图像一部分用于训练，剩余的用于测试，也可以全部用于训练然后选取一部分进行测试。然而，不像人脸识别大则几千万张图片的数据集，行人重识别数据集往往很小，这使得基于数据集的拓展学习变得较为困难。表 2-1 整理了部分常用的行人重识别公开数据集。

表 2-1　行人重识别数据集

数据集名称	身份	摄像头	图片数	标注方式	年份
CUHK03[57]	1467	10	13164	手工/DPM	2014
Market1501[58]	1501	6	32668	手工/DPM	2015
MARS[59]	1261	6	1191003	DPM+GMMCP	2016
DukeMTMC-reID	1812	8	36441	手工	2017

本文实验采用的是 Market1501 数据集，该数据集在清华大学校园中拍摄采集，于 2015 年公开。Market-1501 数据集收集使用了 6 个摄像头。该数据集包含 32,668 个带有 1,501 个身份的带注释的边界框。确保每个带注释的标识出现在至少两个摄像机中，以便可以执行跨摄像机搜索。该数据集中图片的命名如“0002_c3s2_001234_01.jpg”。“0002”表示的是第二个身份的标签，“c3”指的是第三个摄像头，“s2”是该摄像机的第二段摄像，“001234”第 1234 帧（帧速率为 25fps），“01”表示这一帧检测到的第一个边界框。

2.1.2 评价指标

每秒帧数 FPS(Frame Per Second)、精度 Precison 和召回率 Recall 是深度学习中常用的指标。对于目标检测，最常用的度量标准则是平均精度 AP(Average Precison)。然而，AP 通常用于特定类别检测的评价，是基于 Precision-Recall 曲线计算出来。为了得到通用类别检测的效果评估，取各个类别 AP 的平均值得到 mAP(Mean Average Precison)。对于行人重识别算法而言，常用的性价指标则为 mAP、Rank 和 CMC。相关的评价指标计算如下：

① 精度 Precision、召回率 Recall

为了方便地得到精度和召回率的计算公式，引入以下符号：

TP: True Positive，实际为真而被划分为假的个数。

FP: False Positive，实际为假而被划分为真的个数。

TN:True Negative，实际为假而被划分为假的个数。

FN: False Negatve，实际为真而被划分为真的个数。

精度 Precison 的表达式为：

$$\text{Precision}=\frac{TP}{TP+FP} \tag{2-1}$$

召回率 Recall 可用公式表示为：

$$\text{Recall}=\frac{TP}{TP+FN} \tag{2-2}$$

② IOU

IOU(Intersection over Union)[54]，成为重叠比或者交并比。IOU 可以表示候选框与原标记框的重叠率，它介于 0 到 1 之间。IOU 是用来判定对象被正确检测的关键指标。将预测的边界框 bounding box 用符号 b 表示，而实际的边界框为 b^g，则 IOU 可以表示为：

$$IOU(b,b^g)=\frac{area(b\cap b^g)}{area(b\cup b^g)} \tag{2-3}$$

IOU 越高，预测框的位置就越准确。因而在评估时，常常对 IOU 设定一个阈值，如果 IOU 大于这个阈值，则认为该预测为 TP，否则为 FP。该值常常是判定

检测对象分类正确与否的标准，其算法流程如下：

算法

输入：$\{(b_j,p_j)\}_{j=1}^{M}$：图像 I 关于 c 类对象的 M 个预测，按降序排列的置信度 p_j；

$B=\{b_k^g\}_{k=1}^{K}$：图像 I 关于 c 类对象的实际框集合

输出：一个决定每个(b_j,p_j)属于 TP 还是 FP 的二进制向量

执行：

初始化 $a = 0$

for j =1,…,M **do**

设置 $A = \emptyset$ and $t = 0$

foreach 未匹配框b_k^g在集合 B 中 **do**

if $IOU(b_j,b_k^g) \geq \varepsilon$并且$IOU(b_j,b_k^g) > t$ **then**

$A = \{b_k^g\}$;

$t = IOU(b_j,b_k^g)$;

end

end

if $A \neq \emptyset$ **then**

设置 $a(i)$=1,预测(b_j,p_j)为 TP

将 A 中匹配成功的实际框从 B 中移除，$B = B - A$

end

end

③ AP、mAP

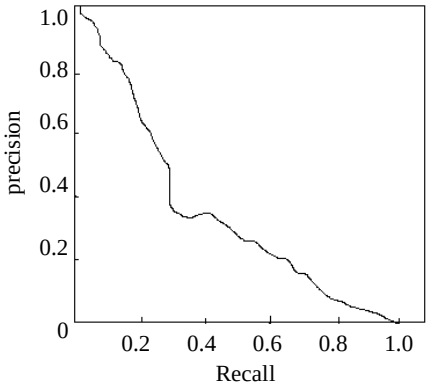

图 2-1 P-R 曲线

图 2-1 展示的是精度-召回率曲线，该曲线是以召回率为横坐标，精度为纵坐

标，统计得到不同阈值下的值并用曲线绘制得到。假设 P-R 曲线可以用函数 P(R)表示，则 AP 的求解即 P-R 曲线与坐标轴围绕的图形面积，而 mAP 则是对 classes 个类别的 AP 求解平均值。

$$AP = \int_0^1 P(R)d(R) \tag{2-4}$$

$$mAP = \frac{1}{classes}\sum_{i=1}^{classes}\int_0^1 P(R)d(R) \tag{2-5}$$

在行人重识别中，mAP 反映的是检索的人在底库中所有正确图片排在结果队列中靠前的程度。假设查询目标在底库中的次序是一组序列$\{q_i|i=1,2,\ldots.n\}$，n 为查询目标在底库中的个数。总共的查询目标有 k 个，则 mAP 可以用公式(2-6)进行简化计算。

$$mAP = \frac{1}{k}\sum\frac{1}{n}\sum_{i=1}^{n}\frac{i}{q_i} \tag{2-6}$$

④ Rank

Rank 是行人重识别的另一个核心评价指标，一般有 Rank1、Rank5、Rank10 等。Rank1 即首位命中率，指的是查询结果第一位的图片为查询目标本身的概率。Rankn 可以反映排在前面的图片的性能，但可能因为存在偶然因素的情况而反映不够全面。

2.2 损失函数

① Softmax 损失函数计算

Softmax 是将网路得到的值进行归一化处理使其值可以解释为类似于概率的过程。因此 Softmax 处理过后的值将在[0,1]的范围之间。下图给出了 Softmax 的处理过程：

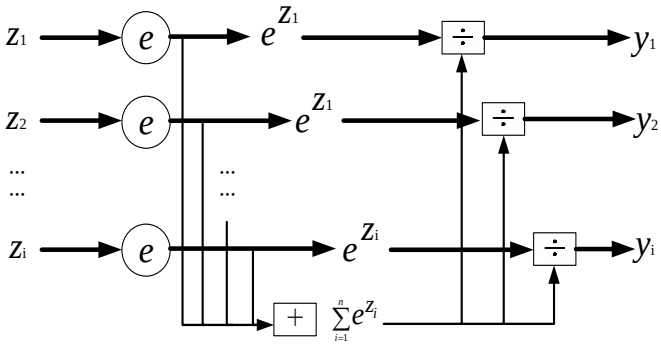

图 2-2　Softmax 处理过程

在神经网络后面添加 Softmax，经过 Softmax 得出的值就是预测的结果。借鉴人脸识别的 Normface loss 可以得到 Softmax 的损失函数：

$$L_{softmax} = -\sum_{i=0}^{N} log \frac{e^{W_{y_i}^T f_i}}{\sum_{k=1}^{C} e^{W_k^T f_i}} \tag{2-7}$$

② Triplet 损失函数计算

Triplet 的核心思想是通过选择三张图片构成一个三元组，即 anchor、negative、positive，通过 triplet loss 学习使得 positive 和 anchor 之间的距离最小，negative 之间的距离最大。其中，anchor 是和 positive 同一类的随机选取的样本，而 negative 和它们属于不同样本。如下图所示：

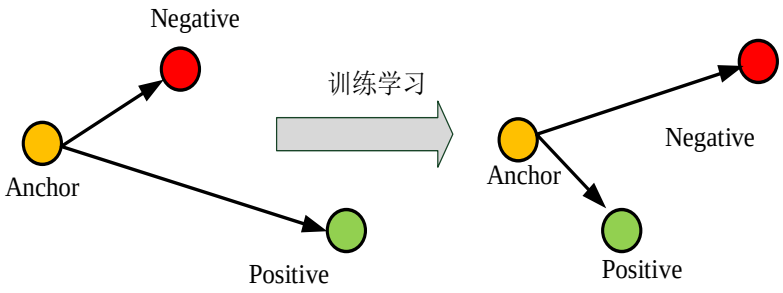

图 2-3　triplet 损失函数学习过程

因此，对于行人重识别问题而言，triplet loss 损失可以较好地解决类间相似、类内差异的问题。也就是说，triplet loss 可以有效处理同一个行人因为场景变换导致的差异。另外，构造三元组是可以缓解因为数据集较小而产生的过拟

合问题。

Triplet Loss 损失函数计算公式如下：

$$L_{triplet} = -\sum_{i=1}^{P}\sum_{a=1}^{K}[\alpha + \max_{p=1,..K}\left\|f_a^i - f_p^i\right\|_2 - \min_{\substack{n=1,..K \\ j=1,...P \\ j\neq i}}\left\|f_a^i - f_n^i\right\|_2] \quad (2\text{-}8)$$

其中，f_a^i和f_p^i属于同一个 ID，f_a^i和f_n^i属于不同的 ID。

2.3 YOLOv3 行人检测算法

本实验采用 yolov3 的目标检测方法来进行行人的检测，作为行人重识别的预处理步骤。

2.3.1 网络结构

	类型	卷积信息		特征图大小
	卷积层	32	3×3	416×416
	卷积层	64	3×3/2	208×208
1×	卷积层	32	1×1	
	卷积层	64	3×3	
	残差层			208×208
	卷积层	128	3×3/2	104×104
2×	卷积层	64	1×1	
	卷积层	128	3×3	
	残差层			104×104
	卷积层	256	3×3/2	52×52
8×	卷积层	128	1×1	
	卷积层	256	3×3	
	残差层			52×52
	卷积层	512	3×3/2	26×26
8×	卷积层	256	1×1	
	卷积层	512	3×3	
	残差层			26×26
	卷积层	1024	3×3/2	13×13
4×	卷积层	512	1×1	
	卷积层	1024	3×3	
	残差层			13×13

图 2-4 Darknet-53 网络结构组成

YOLOv3 采用了新的网络结构用于特征的提取。该网络结构类似于 YOLOv2

的网络 Darknet-19。不同的是，YOLOv3 增加了新的残差网络结构用于辅助。另外，该网络有连续的 3×3 和 1×1 卷积层，并存在一些跨层的连接。相较于 DarketNet-19，该网络结构更大。由于网络中存在 53 个卷积层，该网络结构被命名为 Darknet-53。Darknet-53 的结构组成如图 2-4 所示，其中一些方框左侧的几乘数字表示该残差组件的重复个数。

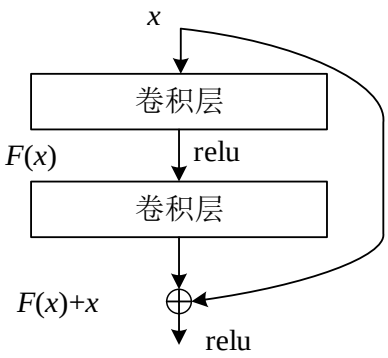

图 2-5　残差层结构

图 2-5 是每个残差层的具体结构，该结构由两个卷积层和一个快捷链路组成。

2.3.2　先验框的提出

YOLOv3 采用了先验框[55]（anchor boxes）。每个网格预先设定一组不同大小和宽高比的边框，来覆盖整个图像的不同位置和多种尺度，这些先验框作为预定义的候选区在神经网络中将检测其中是否存在对象，以及微调边框的位置。先验框的加入大幅度提高了召回率 Recall。

为了减少先验框的个数造成的复杂度，利用聚类算法计算两个边框的“距离”：

$$d(b,centroid)=1-IOU(b,centroid) \tag{2-9}$$

最后聚类出 9 种尺度的先验框。在 COCO 数据集上这 9 个先验框分别是：(10×13)，(16×30)，(33×23)，(30×61)，(62×45)，(59×119)，(116×90)，(156×198)，(373×326)。分配上，较小的特征图上有最大的感受野，应用较大的先验框，适合检测较大的对象；最大的特征图上有最小的感受野，应用较小的先验框，适合检测较小的对象。下表给出了特征图和先验框的选择关系。

表 2-2　特征图与先验框

特征图	13×13	26×26	52×52
先验框	(116×90)	(30×61)	(10×13)
	(156×198)	(62×45)	(16×30)
	(373×326)	(59×119)	(33×23)

2.3.3　边界框的预测与编码

YOLO 的边框中心约束在网格内。假设单元格从图像的左上角偏移(c_x，c_y)并且距离最小的先验框具有宽度 p_w 和高度 p_h，如图显示：

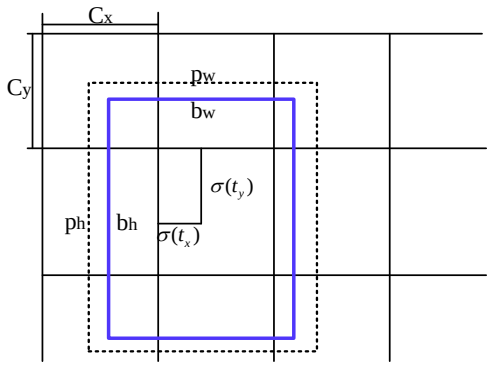

图 2-6　具有优先尺寸和位置预测的边界框

t_x、t_y、t_w、t_h 分别作为预测边框中心和宽高的参数，则预测边框的中心位置(b_x，b_y)和宽高(b_w，b_h)可用以下公式表示：

$$\begin{aligned} b_x &= \sigma(t_x) + c_x \\ b_y &= \sigma(t_y) + c_y \\ b_w &= p_w e^{t_w} \\ b_h &= p_h e^{t_h} \end{aligned} \tag{2-10}$$

若用 $Pr(Object)$表示该边界框存在对象的概率，则该边界框的置信度分数可表示为：

$$Confidence = \Pr(Object) \cdot IOU(b, object) = \sigma(p_0) \tag{2-11}$$

其中，σ 是 sigmoid 函数，这里对预测参数 p_0 进行 σ 变换后作为置信度的值。

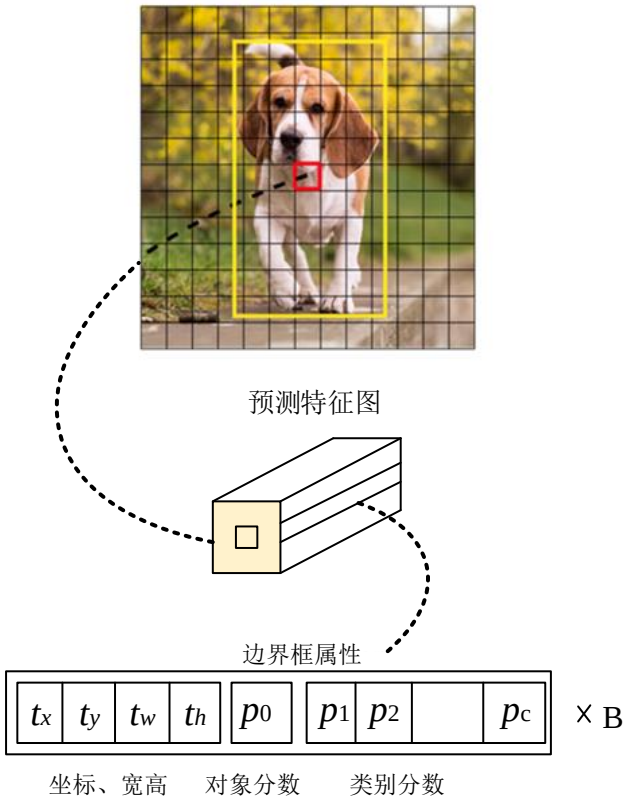

图 2-7　编码输出结构

图 2-7 给出了特征图边界框属性输出编码的结构，它们对应于预测框中心点所在网格。其中，前 5 个属性值分别代表边界框的中心坐标、尺寸、对象置信度分数，后面 C 个属性值代表对 C 个类别的置信度分数。本实验采用具有 80 个类别的 COCO 数据集，因此 C 等于 80。

2.3.4　非极大值抑制

非极大值抑制(NMS, Non-Maximum Supptrssion)，是一种在同一个对象有多个候选框时选择置信度分数最高的一个候选框并消除冗余候选框的策略。图 2-8 展示了该策略的执行策略流程：

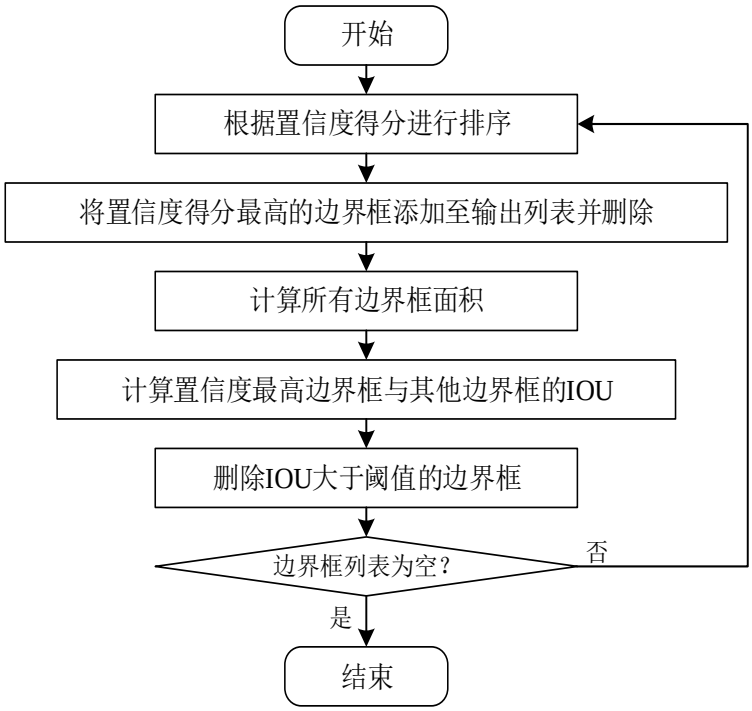

图 2-8　NMS 策略执行过程

2.4　行人重识别算法实现思路与基础网络

行人重识别是判断出现在不同监控视频中的行人是否为同一行人的技术，也被认为是跨境跟踪问题或图像检索问题。行人重识别在 2006 年作为单独的术语被提及[56]。该技术具有很强的实用性和巨大的实际需求，可以帮助手机用户实现相册聚类、帮助零售或为经营者获取有效的顾客轨迹挖掘商业价值，也可以追踪失踪儿童，犯罪分子等。本文 1.2.2 小节综述了解决行人重识别问题的若干种方法，较为常用的是基于特征学习和度量学习的行人重识别方法。

行人重识别问题包括基于图像和基于视频两种类型，由于目前行人重识别数据集设计多为基于图像的形式，因此大多数研究者关注基于图像的行人重识别。在基于图像的行人重识别问题中，常常需要先在视频中进行行人检测定位行人的位置。然后，将检测到的行人放入图片底库作为行人识别的查询图集。

行人重识别的实现思路大致可以分为两个阶段：第一个阶段是将查询目标行人的图像和底库中全部行人的图像分别通过卷积网络进行特征提取，提取的特征一般抽象为向量形式，第二阶段为将检索图与底库的特征进行距离计算(例如欧

式距离)，将这些距离从小到大排序（或者转换成相似度得分进行倒序排列），排名越靠前则表示相似性越大。ReID 的实现思路也决定了研究的两个主要方向：一个是通过改进网络提取更好的特征表示，不同的研究机构会设计不同的网络结构来提取特征，在训练时，会设计损失函数并最小化损失使得训练得到的特征更加有意义；另一个是探究高效的相似性度量方法计算比对不同特征之间的距离。

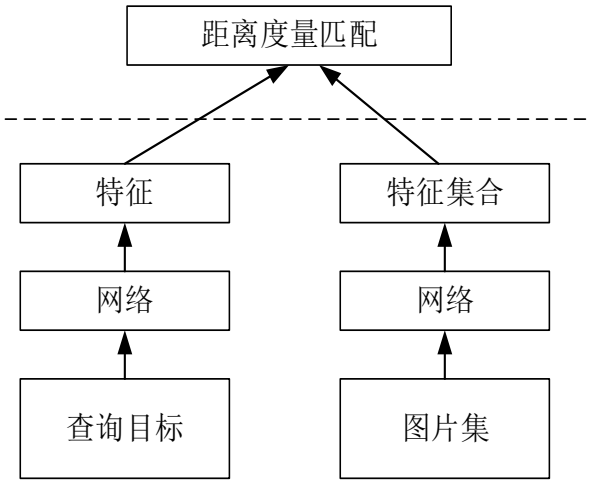

图 2-9　ReID 实现思路框架

本文主要研究如何通过更好的网络提取到更全面的特征。现有的 ReID 研究很多都是基于 Resnet50[54]的网络结构，因为它表现很好而且结构简洁。Resnet50 一般分为 5 层，图像输入为(224*224*3)，每层输出的特征图谱长宽都会比上一层缩小一半。然后进行池化，所谓池化就是每个特征图谱里取一个最大值或者平均值。最后基于这个特征进行分类。其简化结构如图 2-10。

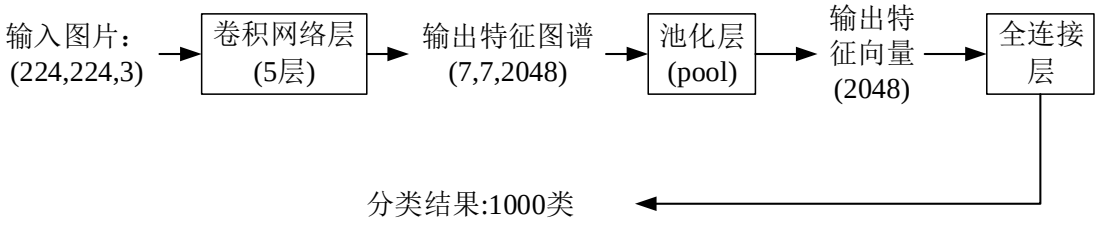

图 2-10　Resnet50 结构

本实验采用的基础网络结构为 PCB 结构[60]，其骨干网络仍旧采用 Resnet50。PCB 的网络结构在原始文章中结构如下图：

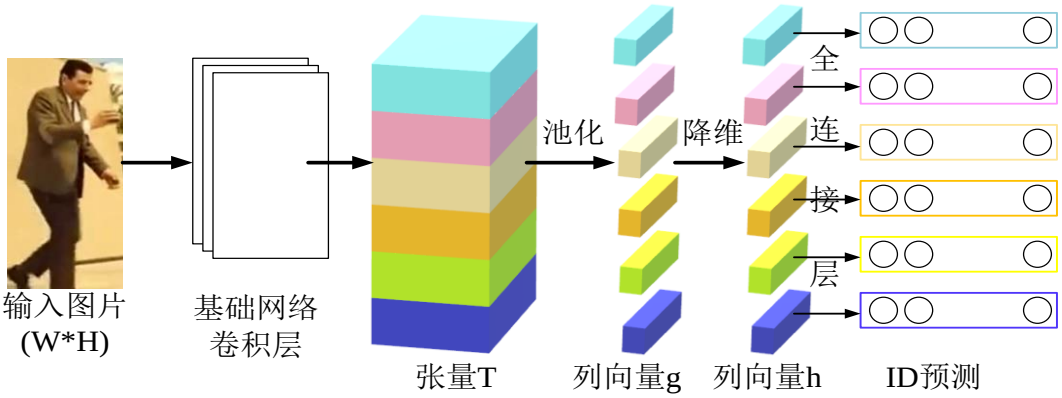

图 2-11　PCB 网络结构

在全局平均池化 GAP 层（global average pooling layer）之前的结构和 Resnet50 一样，而 GAP 层进行了一定的改变并移除了池化层之后的所有层。PCB 的具体实现过程为：输入一张图片，经过骨干网络生成激活的三维深度特征张量 T。张量 T 中定义沿着通道轴观察的激活向量为列向量 f。然后，将张量 T 平均分成 p 个水平块，分别对每个水平块进行平均值池化得到 p 个全局特征 g。之后采用卷积层降低 g 的尺寸得到列向量 h。最后，将每个 h 输入到分类器中，利用完全连接层和随后的 Softmax 函数实现，以预测输入行人的身份。

2.5　本章小结

本章主要讲述行人检测和行人重识别的相关知识。2.1 节介绍了行人重识别需要使用的数据集以及目标检测和行人重识别的性能评价指标。2.2 节给出了行人重识别中的分类损失函数 Softmax Loss 以及度量损失函数 Triplet Loss，作为衡量模型预测好坏的工具。特别地，2.3 节详细介绍了本实验采用的 YOLOv3 算法，该小节分别从网络架构、先验框的生成、边界框的预测与编码、非极大值抑制流程进行阐述。最后，2.4 节分析了行人重识别算法问题解决的思路，提出本文的研究重点是特征提取的网络结构设计，给出了 ReID 算法常用的基础网络 ResNet50 和 PCB 的网络结构。

第 3 章　行人识别的实验设计

3.1　实验基础库和参数

3.1.1　基础库

① Pytorch

本算法测试实验基于 Pytorch 的深度学习框架，语言实现较为简单灵活。因为 Pytorch 强大的 GPU 加速张量计算和包含自动求导系统（autograd）的深度学习网络而迅速受到 AI 研究人员的推崇。

② Torchvision

Torchvision 是独立于 pytorch 的关于图像操作的工具库。在本实验中主要应用到了该库中的几个重要的软件包：

- vision.datasets :本实验在进行目标检测时使用了该库中的 COCO 数据集。
- vision.models: 主流的深度学习模型，本实验在进行行人重识别时导入了 ResNet50 和 PCB 网络模型。
- vision.transforms : 常用的图像操作，在本实验中主要应用了图像、numpy 数组和张量 tensor 之间的转化以及一些其他操作。

3.1.2　调整参数

在本实验的算法训练过程中，对于行人重识别算法而言，有许多的参数，很大程度上直接影响了训练结果的优劣。但是，目前并没有足够的理论依据来完成最优的参数设置，因此有必要进行多次尝试。

① Batchsize

Batchsize 是每次送入网络训练的数据的样本数量，为了在内存效率和内存容量之间寻求最佳平衡，batchsize 应该精心设置，从而最优化模型的性能和速度。

② Learning Rate

学习率是深度学习最重要的参数之一，决定着目标函数能否收敛到局部最小值以及何时收敛到最小值。若学习率过小，则收敛过程将十分缓慢；若学习率设

置过大，则梯度值将在最小值附近来回震荡甚至可能无法收敛。

③ Learning Rate Decay

如果固定学习率，当达到收敛状态时容易在最优值附近某个较大的区域内摆动，因此随着迭代次数的增加，有必要减小学习率，使得收敛时摆动的区域较小。本实验采用离散轮次下降法，即训练次数达到固定次数的 epoch 后，将学习率下降为原来的 10%，但何时进行学习率的衰减，也需要多次实验尝试。

3.2 实验流程设计

本实验采用行人检测+行人重识别的模式来进行人物的识别。实现过程如下：

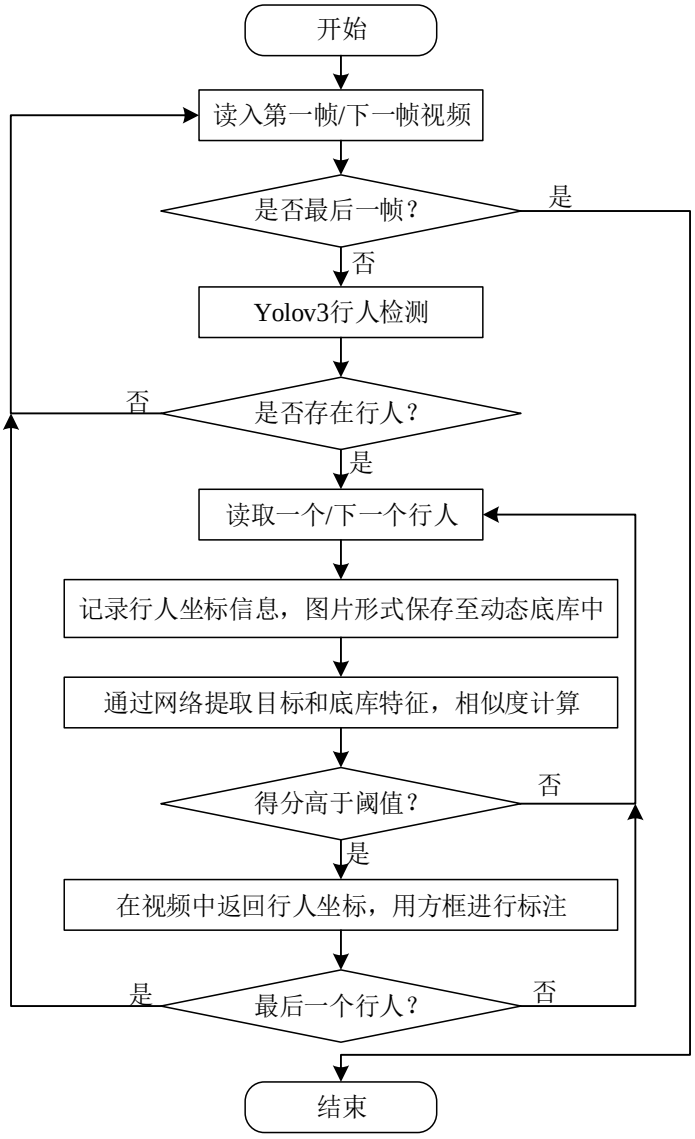

图 3-1 实验流程

本实验首先读取视频，判断是否为视频最后一帧，如果是则结束本实验，否则进一步进行目标检测检测出该视频帧中的所有行人。若没有检测到一个行人，则跳到下一帧进行检测，若检测到多个行人，则保存这些行人坐标信息，并将行人以方框的形式截取成图片，保存至视频底库中。然后对底库中的行人图片和查询目标图片分别进行特征提取并进行相似度计算，即“距离”计算或相似度得分计算。距离越小、相似度得分越高、匹配成功的概率也越高。对相似性得分设定一个阈值，高于该阈值则认为检测到的行人为查询目标的嫌疑较大，应该在视频中进行标注，以便帮助警务工作的进行。基于现实生活中警务需求巨大的实际，如果每检测到一个行人并将其保存为图片，则将大大占用系统资源，因此，在对该帧进行相似度匹配后，本实验将动态地删除这些图片，释放系统资源。由于实验从目标检测到行人重识别，都是端到端的学习方式，因此实验对于实时性的检测将同样适用。

3.3 多粒度网络设计

以 ResNet50 网络结构为基网络，受 PCB 分层提取局部特征的启发，设计了一种兼顾全局与局部特征的多粒度网络。

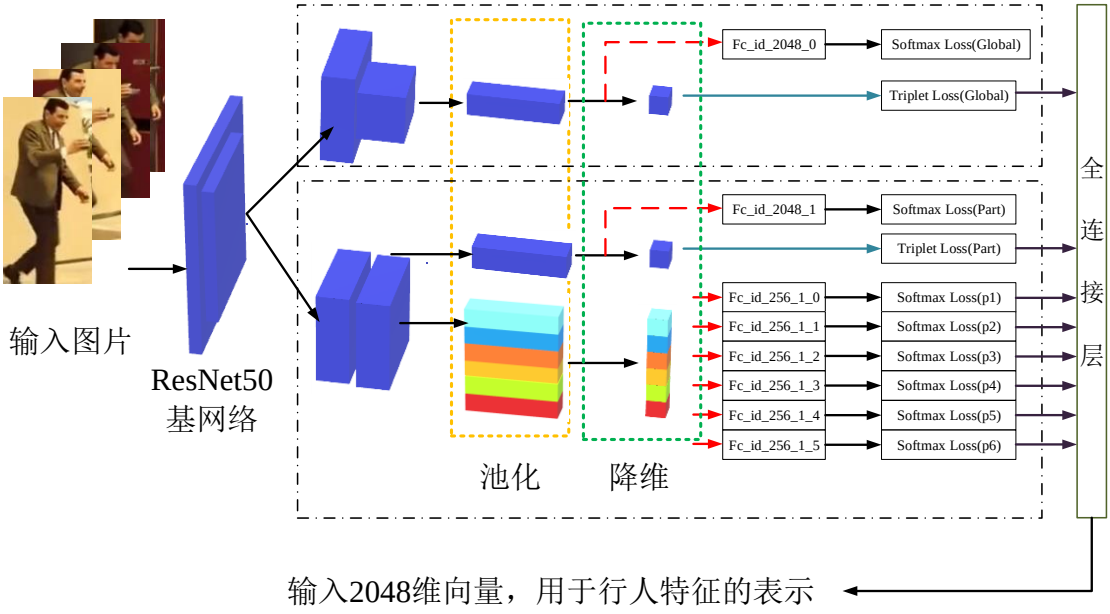

图 3-2　多粒度网络结构

如图 3-2 所示，网络结构从左到右先经过 2 个模块，这 2 个模块代表的是 2 个分支共享网络，这两个分支网络在前三层是共享的，到第四层分成两个支路。

上面的支路用于全局特征提取，先在 res_conv5_1 中使用了 stride 等于 2 的卷积进行下采样，然后对得到的特征图谱采用全局最大池化生成 2048 维的特征向量，并通过 1×1 的卷积压缩为 256 维的特征向量。

下面的支路用于局部特征的提取，为了保留适合局部特征的 ROI，该分支没有使用 stride 等于 2 的卷积下采样。然后在纵向上均匀地将特征图谱划分为 6 层，之所以划分成 6 层是因为 PCB 网络结构通过实验发现 6 层的分层提取细节特征更好。最后，对每层分别进行全局最大池化和 1×1 卷积降维来得到对应的局部特征。

在网络的最后可以得到 2 个 256 维向量的全局特征向量和 6 个 256 维局部特征向量，经过全连接层得到 2048 维向量用作行人的特征表示。本实验将分别对目标和搜索底库中的行人图像通过该网络进行特征的提取，并通过提取特征的相似性度量，给出与目标相似的底库中的行人图像。

3.4 本章小结

本章主要阐述行人识别实验的设计。3.1 节给出了实验设计的基础语言框架和工具库，并就实验需要考虑的参数设置进行了说明。3.2 节给出了整个实验的实现流程框架图并进行了说明。3.3 节以 ResNet 为基网络，结合 PCB 的局部特征提取思想，设计了一种兼顾全局特征和局部特征的多粒度网络。

第 4 章　基于多粒度网络的人物识别算法测试与实现

4.1　Yolov3 行人检测算法实现

YOLOv3 是通用的目标检测算法，可以检测近 80 种类别的物体，因此需要通过判断仅识别行人这一类。yolov3 在 github 上公开了代码，需要在此基础上加入是否为行人的判断语句，在输出识别类别时只对 Person 类进行标注。图 4-1 给出了行人检测的结果。

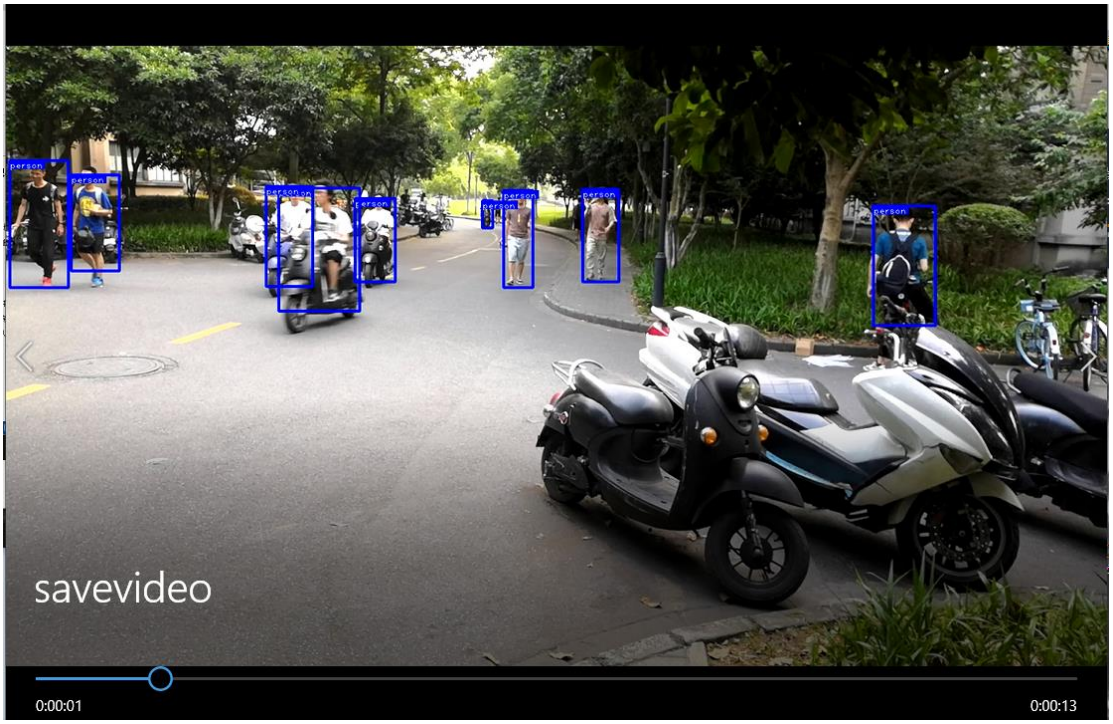

图 4-1　行人检测结果

可以静态地将视频中每帧检测到的行人以方框的形式进行定位展示，并以方框为界将行人裁剪为 jpg 图片保存至视频查询底库中，图 4-2 给出了查询底库的静态建立，检测到的每个人将以 Market1501 数据集的命名格式存入图片底库中，作为警务工作中查询目标人物的搜索图片库。

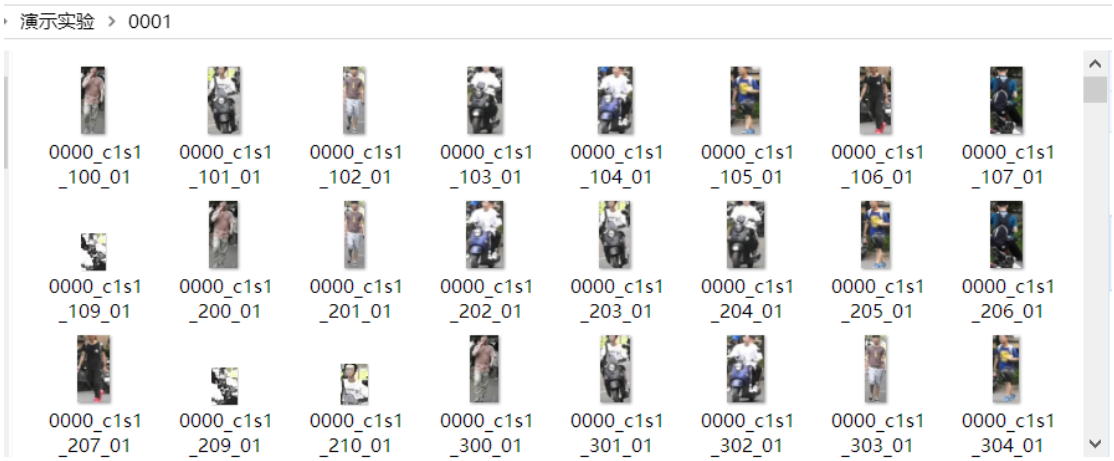

图 4-2　静态查询底库的建立

4.2　行人重识别算法测试

4.2.1　基于深度学习的基础网络测试

本实验设计的多粒度网络的基础结构采用 PCB 网络结构，在 Market1501 数据集上进行训练与测试。用 mAP 和 Rank1 指标进行评价。

4.2.1.1　过程监督

为反映训练过程的好坏，绘制 loss 曲线和 top1err 曲线，观察在测试集上 loss 的变化趋势。

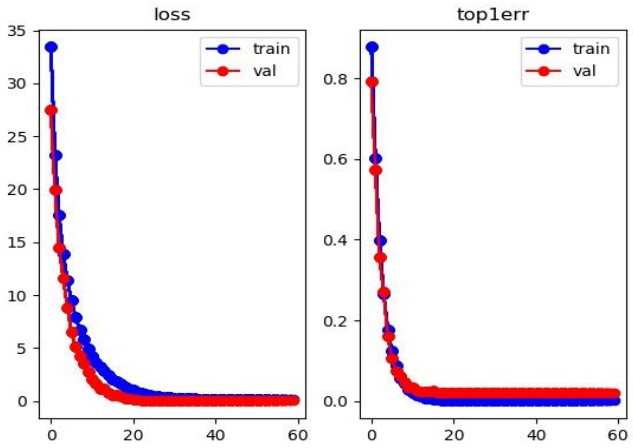

图 4-3　训练过程曲线

图 4-3 显示，在测试集上 val loss 随着 epoch 的增大而减小，因此训练过程未

发生过拟合的情况。另外，误差曲线也逐渐下降并不断趋向于 0，这表明分类的精度 acc 在不断增加并收敛趋向于 1。Val-loss 可以采用 softmax loss、triplet loss 或者其他损失函数，但其变化趋势可以反映过拟合的情况，并给参数的优化调整提供依据。

4.2.1.2 调参优化

由于深度学习包含许多参数，这些参数的选取对于结果有很大的影响。因此，有必要进行多组对照实验，以选取相对最优的参数。在测试机上发生过拟合时，也应该调整参数。评价该组参数设置下模型的好坏仍用评价指标 Rank1 和 mAP 进行表示。针对 PCB 网络模型，给出 batchsize、leaning rate 等参数的调参过程。

① Part

PCB 网络结构中将特征图谱纵向平均分割成 6 部分，为了验证分层数对局部特征提取的影响，设置分层数分别为 1(不分层)、2、4、6、8、12。当 p 为 1 时，提取的为全局特征。

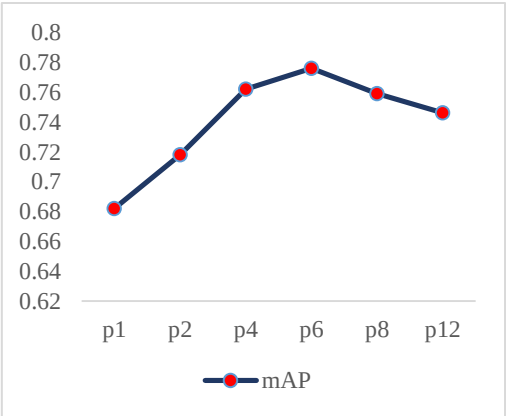

图 4-4(a) part 对训练 mAP 的影响

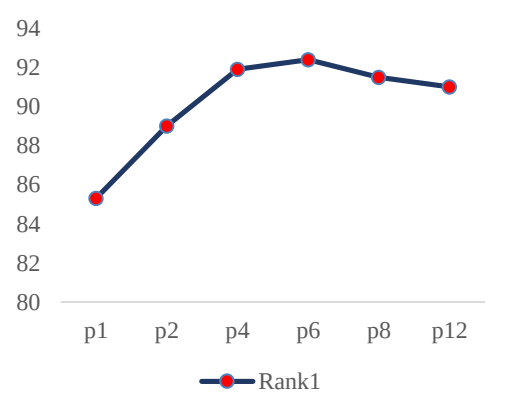

图 4-4(b) part 对训练 Rank1 的影响

实验显示，分层数一开始的增加会导致细粒度程度增大，有利于局部特征的提取，当 p 超过 6 后，mAP 和 Rank1 开始下降。因此，并不是分层数越大对局部特征的提取越有利，当分层数过大时，很有可能导致人的特殊部位被分割，会不利于局部特征的提取表示。

② Batchsize

分别取 batchsize 为 4、8、16、32，其余参数设置相同，计算 Rank1 和 mAP，进行比较。

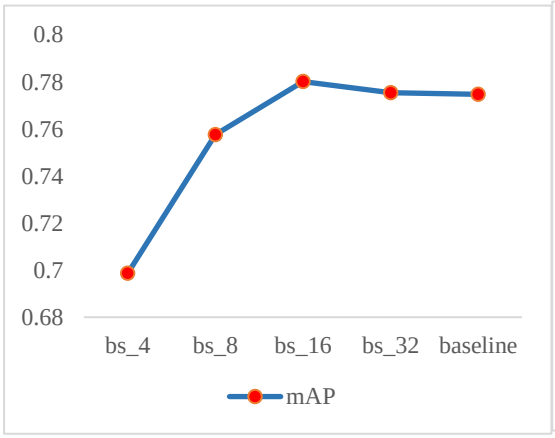

图 4-5(a)　batchsize 对训练 mAP 的影响

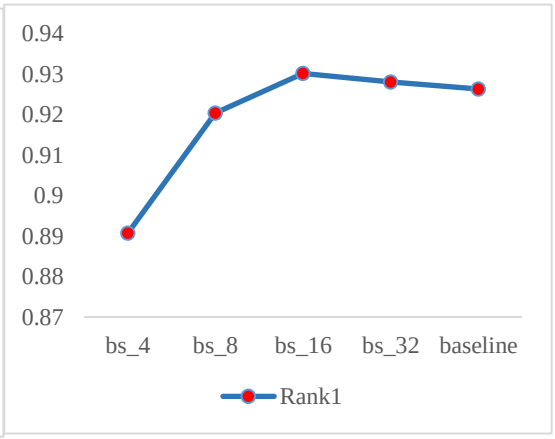

图 4-5(b)　batchsize 对训练 Rank1 的影响

实验显示，对 PCB 模型而言，batchsize 大小选取 16 训练效果最佳。

③ Learning rate

取 batchsize 为 16，learning rate 分别设置为 0.005、0.01、0.02、0.04、0.08，其余参数设置相同，计算 Rank1、Rank5、Rank10、mAP。

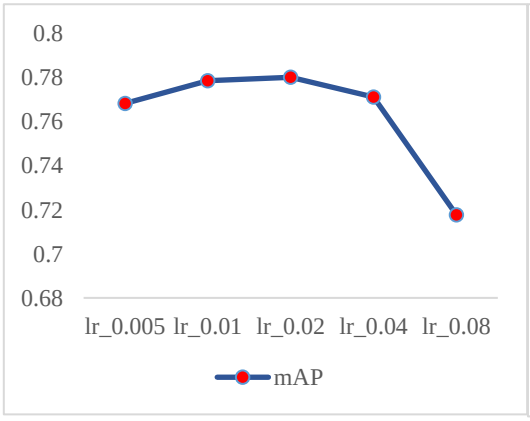

图 4-6(a)　learning rate 对训练 mAP 的影响

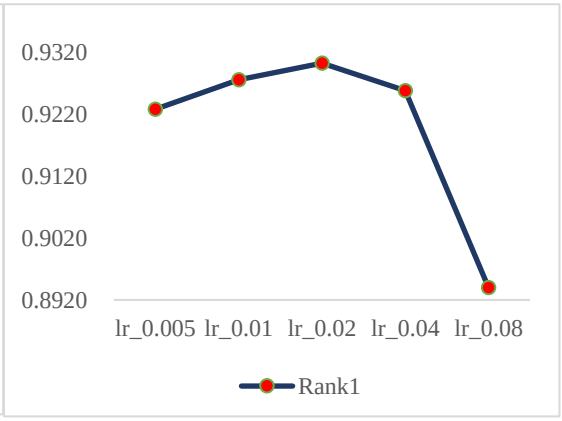

图 4-6(b)　learning rate 对训练 Rank1 的影响

实验显示，Learnig Rate 取值范围在 0.01-0.04 之间模型训练效果较好。当然，learning rate 的选取依赖于网络结构，但可以得到实验结论是，过大的 learning rate

容易产生震荡，导致训练和测试的性能不佳。

4.2.2 基于深度学习的多粒度网络测试

上一节给出了基础网络 PCB 在 Market1501 数据集上的测试，并给出了相关参数的优化过程。可以确定的是，分层数为 6 时的 Rank1 和 mAP 指标最高，这表明将图像分为 6 层提取局部特征效果较好。另外，实验也显示，batchsize、learning rate 等参数的设置对训练结果有较大的影响。

以 PCB 网络的训练测试调参过程为参考，设置基于多粒度网络的参数：batchsize 设置为 16，learning rate 设置为 2e-4，epoch 设置为 400，在 epoch 为 320 和 380 的时候降低学习率为之前学习率的 10%。在网络结构中，分别对全局特征和局部特征采用 triplet loss 和 softmax loss (crossentropy loss)。给出测试过程中 loss 曲线的变化趋势。

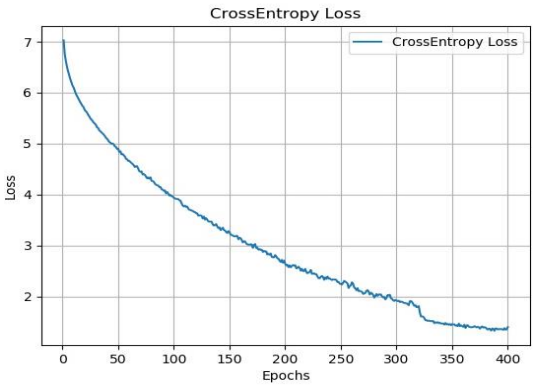

图 4-7(a) softmax loss 变化曲线

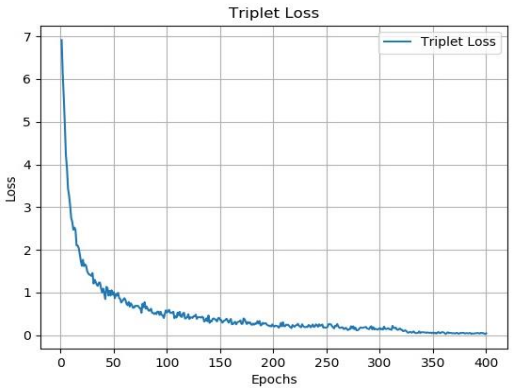

图 4-7(b) triplet loss 变化曲线

可以得到，loss 值随着 epoch 的增大逐渐下降，因此训练未发生过拟合的情况。特别地，triplet loss 的曲线下降较快，而 total loss 在 epoch 为 320 的学习率下降后已经趋向于收敛，因此，可以认为多粒度网络的训练结果已经达到最优点附近，至于是否为最优结果，则需要大量的时间进行多组实验得到。同样地，计算 mAP 和 Rank1。

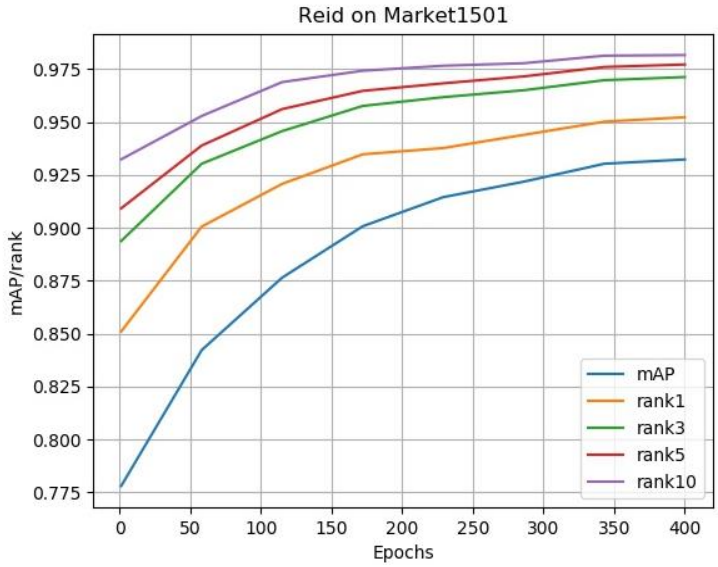

图 4-8　多粒度网络模型测试性能指标

图 4-8 显示，多粒度网络在 Market1501 数据集上测试得到了 93%左右的 mAP 和 95%以上的 Rank1。

4.2.3　网络模型的性能比较

表 4-1　网络模型在 Market 数据集上的测试结果比较

model	Rank1	Rank5	Rank10	mAP
Resnet50	0.889549	0.964667	0.975653	0.732591
ResNet50_paper	0.8884	--	--	0.7159
PCB	0.930226	0.973575	0.982779	0.780155
PCB_paper	0.9264	--	--	0.7747
多粒度网络	0.9522	0.9712	0.9816	0.9323

上表给出了 ResNet50 和 PCB 模型发表论文中基于 Market1501 数据集训练测试的 Rank 值和 mAP 值，分别在用模型名+“paper”的形式进行表示。本实验对此进行复现并在多组调参实验后，本人得到的训练结果指标均优于原文指标。而本实验设计的多粒度网络的 Rank1 值比 PCB 提高了 2 个百分点，mAP 值则提高了近 15 个百分点。

4.3 行人识别结果展示

通过 python 命令输入查询对象的编号，在采集得到的视频流中识别该人物。对识别结果中的几帧进行截图，展示如下：

图 4-9 行人识别的效果展示

4.4 本章小结

本章是基于多粒度网络的人物识别算法测试与实现。4.1 节给出了 yolov3 行人检测的结果，并展示了静态查询图集的生成结果。4.2 节是行人重识别算法的测试，该节给出为了防止训练过拟合提出的过程监督，通过 val loss 的变化趋势判断训练是否过拟合。以 PCB 网络为基础网络给出参数调整优化的过程。然后针对本文提出的多粒度模型进行测试，分别记录了 Softmax 损失和 Triplet 损失的变化过程，并给出了 mAP 和 Rank1 随着 epoch 的增大的变化趋势。最后 4.3 节展示监控视角下对某个行人的识别效果展示。

第 5 章　总结与展望

5.1　本文工作总结

行人重识别是实现行人检索与跟踪问题的关键技术，是在人脸识别无法有效进行时可以替代的方法。目前，基于警务工作的行人检索有巨大的市场需求。因此，对行人重识别的研究具有重要的现实意义。行人重识别方法主要涉及特征提取和距离度量两个方面，本文重点关注如何提取有效的特征。

本文首先简述了本课题的研究背景与意义，综述了基于深度学习的人物识别方法。然后，介绍了基于深度学习的目标检测的相关知识，给出了 yolov3 用于目标检测的详细过程，作为行人重识别的预处理步骤。接下去，阐述了行人重识别的数据集、性能评价指标、网络模型，并进行相应的实验。下面对本文的主要工作进行简要总结说明：

(1) 本文综述了基于深度学习的目标检测方法，主要研究了基于端到端学习的 yolov3 算法实现的原理，有效进行了行人检测，包括部分受到遮挡的人物，并将检测到的人物保存入库制作查询图集。

(2) 本文就行人重识别算法基于基础网络 ResNet50 和 PCB 网络结构进行复现并通过调参优化实现了优于原文的性能指标，并以 PCB 提取局部特征为启发，设计了一种兼顾全局特征和局部特征提取的多粒度网络，在 Market1501 数据集上测试得到了 93%左右的 mAP 和 95%左右的 Rank1

(3) 本文结合行人检测和行人重识别，实现了监控视角下对某个特定人物的识别。

5.2　未来工作展望

本文虽然实现了目标检测和行人重识别的算法，并基于路口拍摄的短视频实现了较为满意的结果，但离实际可应用的程度还有很大差距。

(1) 现实生活中的情景非常复杂。实际监控下的行人多有遮挡的情况，目前的目标检测算法无法有百分百的识别精度，无法保证查询目标全部被目标检测算

法检测出来并保存在查询图库中。本实验中，出现了将电动车识别为行人的情况，降低检测阈值可以避免但是也会导致部分行人无法被检测到。因此，设计一个精度更高的目标检测算法是实现行人识别工作的重要保障。

(2) 行人重识别的数据集较少，目前主流的行人重识别数据集都是直接给定的人物，而且图片大小统一，行人在图片中占比较大。但监控视角下无法固定行人大小的统一，随着行人与监控摄像头距离的变化，行人在视频中的占比也将不断变化，本实验中明显出现了视频远端的行人无法有效地被识别的情况。因此建立一个更大、更全面的数据集至关重要。

(3) 参数的优化调整是一个极其耗时的工作，因此建立一套参数优化调整策略变得极其重要。

(4) 随着网络层数的增加，深度学习网络训练时长显著变得很长，而且对设备的要求变得更大，更新设备以加快训练又需要很大的花费，这对学生的研究工作的开展造成较大的不便。这也是本文在现有实验条件下对新提出的多粒度网络无法开展多组调参实验的原因。如果时间和设备都允许的情况下，应该尽量多地进行训练测试，并在其他数据集上检验网络性能。

参 考 文 献

[1] G. E. Hinton, S. Osindero, Y.-W. Teh. A Fast Learning Algorithm for Deep Belief Nets[J]. Neural Computation, 2006, 18(7):1527-1554.

[2] G. Marcus. Deep learning: A critical appraisal[J]. ArXiv preprint, 2018, arXiv: 1801.00631.

[3] K. Zhang, Z. Zhang, Z. Li, Y. Qiao. Joint face detection and alignment using multitask cascaded convolutional networks[J]. In IEEE Signal Processing Letters, 2016, 23(10):1499–1503.

[4] Y. Taigman, M. Yang, M. Ranzato, L. Wolf. Deepface: Closing the gap to human-level performance in face verification[C]// In IEEE conference on computer vision and pattern recognition, 2014, pp. 1701–1708.

[5] Y. Sun, X. Wang, X. Tang. Deeply learned face representations are sparse, selective, and robust[J]. In IEEE conference on computer vision and pattern recognition, 2015, pp. 2892–2900.

[6] Y. Sun, D. Liang, X. Wang, X. Tang. DeepID3: Face recognition with very deep neural networks[J]. ArXiv preprint, 2015, arXiv:1502.00873.

[7] Y. Sun, X. Wang, X. Tang. Deep learning face representation from predicting 10,000 classes[C]// In IEEE conference on computer vision and pattern recognition, 2014, pp. 1891–1898.

[8] Y. Sun, Y. Chen, X. Wang, X. Tang. Deep learning face representation by joint identification-verification[J]. In Neural Information Processing Systems, 2014, pp. 1988–1996.

[9] Y. Wen, K. Zhang, Z. Li, Y. Qiao. A discriminative feature learning approach for deep face recognition[M]. In European Conference on Computer Vision - ECCV 2016. Springer International Publishing, 2016, pp. 499–515.

[10] F. Schroff, D. Kalenichenko, J. Philbin. Facenet: A unified embedding for face

recognition and clustering[J]. In IEEE conference on computer vision and pattern recognition, 2015, pp. 815-823.

[11] K. Cao, Y. Rong, C. Li, X. Tang, C. C. Loy. Pose-robust face recognition via deep residual equivariant mapping[C]// In IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 5187–5196.

[12] Y. Lin, L. Zheng, Z. Zheng, Y. Wu, Y. Yang. Improving person re-identification by attribute and identity learning[J]. ArXiv preprint, 2017, arXiv:1703.07220.

[13] H. Liu, J. Feng, M. Qi, J. Jiang, S. Yan. End-to-end comparative attention networks for person re-identification[J]. In IEEE Transactions on Image Processing , 2017, 26(7):3492–3506.

[14] R. R. Varior, M. Haloi, G. Wang. Gated siamese convolutional neural network architecture for human re-identification[J]. In European Conference on Computer Vision. Springer, 2016, pp. 791–808.

[15] D. Cheng, Y. Gong, S. Zhou, J. Wang, N. Zheng. Person reidentification by multi-channel parts-based cnn with improved triplet loss function[C]// In IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1335–1344.

[16] H. Zhao, M. Tian, S. Sun, J. Shao, J. Yan, S. Yi, X. Wang, X. Tang. Spindle net: Person re-identification with human body region guided feature decomposition and fusion[C]// In Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on. IEEE, 2017, pp. 907–915.

[17] S. Yang, P. Luo, C. C. Loy, X. Tang. From facial parts responses to face detection: A deep learning approach[J]. In IEEE International Conference on Computer Vision, 2015, pp. 3676-3684.

[18] H. Li, Z. Lin, X. Shen, J. Brandt, G. Hua. A convolutional neural network cascade for face detection[C]// In IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 5325-5334.

[19] Z. Zhang, P. Luo, C. C. Loy, X. Tang. Facial landmark detection by deep multi-task learning[C]// In European Conference on Computer Vision, 2014, pp. 94-108.

[20] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li∗ , W. Liu∗. CosFace: Large margin cosine loss for deep face recognition[J]. In Computer Vision and Pattern Recognition, 2018.

[21] W. Liu, Z. Li, and X. Tang. Spatio-temporal embedding for statistical face recognition from video[J]. In European Conference on Computer Vision, 2006.

[22] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song. SphereFace: Deep Hypersphere embedding for face recognition [J]. In Computer Vision and Pattern Recognition, 2017.

[23] W. Liu, Y. Wen, Z. Yu, M. Yang. Large-Margin softmax loss for convolutional neural networks[J]. In International Conference on Machine Learning, 2016.

[24] K. Shlizerman, S. M. Seitz, D. Miller, E. Brossard. The megaface benchmark: 1 million faces for recognition at scale[J]. In Computer Vision and Pattern Recognition, 2016, pp. 4873-4882.

[25] L. Wolf, T. Hassner, I. Maoz. Face recognition in unconstrained videos with matched background similarity[C]// In IEEE Conference on Computer Vision and Pattern Recognition, 2011.

[26] W. Chen, X. Chen, J. Zhang, K. Huang. Beyond triplet loss: a deep quadruplet network for person re-identification[J]. ArXiv preprint, 2017,arXiv: 1704.01719v1.

[27] H. Alexander, B. Lucas, L. Bastian. In defense of the triplet loss for person reidentification[J]. ArXiv preprint, 2017, arXiv:1703.07737.

[28] Q. Xiao, H. Luo, C. Zhang. Margin sample mining loss: A deep learning based method for person re-identification [J]. ArXiv preprint, 2017, arXiv: 1710.00478.

[29] R. R. Varior, B. Shuai, J. Lu, D. Xu, G. Wang. A siamese long short-term memory architecture for human re-identification[J]. In European Conference on Computer Vision. Springer, 2016, pp.135–153.

[30] L. Zheng, Y. Huang, H. Lu, Y. Yang. Pose invariant embedding for deep person reidentification[J]. ArXiv preprint, 2017, arXiv:1701.07732.

[31] L. Wei, S. Zhang, H. Yao, W. Gao, Q. Tian. Glad: Global-local-alignment descriptor for pedestrian retrieval[J]. ArXiv preprint, 2017, arXiv:1709.04329.

[32] H. Liu, Z. Jie, J. Karlekar, M. Qi, J. Jiang, S. Yan, J. Feng. Video based person re-identification with accumulative motion context[J]. In IEEE Transactions on Circuits and Systems for Video Technology, 2017.

[33] G. Song, B. Leng, Y. Liu, et al. Region-based quality estimation network for large-scale person re-identification[J]. ArXiv preprint, 2017, arXiv:1711.08766.

[34] Z. Zheng, L. Zheng, Y. Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro[J]. ArXiv preprint, 2017, arXiv: 1701.07717.

[35] Z. Zhong, L. Zheng, Z. Zheng, et al. Camera style adaptation for person re-identification[J]. ArXiv preprint, 2017, arXiv:1711.10295.

[36] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. Berg, F. Li. ImageNet large scale visual recognition challenge[J]. In International Journal of Computer Vision, 2015, pp. 115(3): 211–252.

[37] L. Liu, W. Ouyang, X. Wang, P. Fieguth, J. Chen, X. Liu, M. Pietikäinen. Deep learning for generic object detection: A survey[J]. ArXiv preprint, 2018, arxiv:1809.02165.

[38] A. Krizhevsky, I. Sutskever, G. Hinton. ImageNet classification with deep convolutional neural networks[C]// In Neural Information Processing Systems, 2012, pp. 1097–1105.

[39] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. Berg, F. Li. ImageNet large scale visual recognition challenge[J]. In International Journal of Computer Vision, 2015, 115(3):211–252.

[40] R. Girshick, J. Donahue, T. Darrell, J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation[C]// In IEEE Conference on Computer Vision and Pattern Recognition. 2014, pp. 580–587

[41] R. Girshick. Fast R-CNN[C]// In IEEE International Conference on Computer Vision, 2015, pp. 1440–1448

[42] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, A. Torralba. Object detectors emerge in deep scene CNNs[J]. In International Conference on Learning Representations, 2015.

[43] S. Ren, K. He, R. Girshick, J. Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks.[J]. In IEEE Transactions on Pattern Analysis & Machine Intelligence, 2017, 39(6):1137-1149.

[44] J. Long, E. Shelhamer, T. Darrell. Fully Convolutional Networks for Semantic Segmentation[J]. In IEEE Transactions on Pattern Analysis & Machine Intelligence, 2014, 39(4):640-651.

[45] J. Redmon, S. Divvala, R. Girshick, A. Farhadi. You only look once: Unified, real-time object detection[J]. In Computer Vision and Pattern Recognition, 2016, pp. 779–78.

[46] J. Redmon and A. Farhadi. Yolo9000: Better, faster, stronger[C]// In IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 6517–6525

[47] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich. Going deeper with convolutions[C]// In IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1–9.

[48] K. He, X. Zhang, S. Ren, J. Sun. Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification[J]. In International Conference on Computer Vision, 2015, pp. 1026–1034.

[49] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Fu, A. Berg. SSD: single shot multibox detector[J]. In European Conference on Computer Vision, 2016, pp. 21–37.

[50] M. Everingham, L. V. Gool, C. Williams, J. Winn, A. Zisserman. The pascal visual object classes (voc) challenge[J]. In International Journal of Computer Vision, 2010, pp. 88(2):303–338

[51] M. Everingham, S. Eslami, L. V. Gool, C. Williams, J. Winn, A. Zisserman. The pascal visual object classes challenge: A retrospective[J]. In International Journal of Computer Vision, 2015, 111(1): 98–136

[52] J. Deng, W. Dong, R. Socher, L. Li, K. Li, F. Li. ImageNet: A large scale hierarchical image database[J]. In Conference on Computer Vision and Pattern Recognition, 2009, pp. 248–255

[53] T. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, L. Zitnick. Microsoft COCO: Common objects in context[J]. In European Conference on Computer Vision, 2014, pp. 740–755

[54] K. He, X. Zhang, S. Ren, J. Sun. Deep residual learning for image recognition[J]. In Computer Vision and Pattern Recognition, 2016, pp. 770–778

[55] N. Gheissari, T. B. Sebastian, and R. Hartley. Person reidentification using spatiotemporal apperance[J]. In 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recogntion, 2006, pp.1528-1535.

[56] W. Li, R. Zhao, T. Xiao, X. Wang. Deepreid: Deep filter pairing neural network for person re-identification[C]// In IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 152-159

[57] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, Q. Tian. Scalable person re-identification: A benchmark[C]// In IEEE International Conference on Computer Vision. IEEE Computer Society, 2015, pp. 1116-1124.

[58] L. Zheng, Z. Bie, Y. Sun, J. Wang, C. Su, S. Wang, Q. Tian. Mars: A video benchmark for large-scale person re-identification[M]// In European Conference on Computer Vision. Springer International Publishing, 2016, pp. 868-884.

[59] Y. Sun, L. Zheng, Y. Yang, Q. Tian, S. Wang, et.al. Beyond part models: person retrieval with refined part pooling[J]. ArXiv preprint, 2018, arXiv:1711.09349v3.

致 谢

四年的大学生活转眼过去，我也即将结束本科毕设。不久之后，我将不得不和我亲爱的母校暂别了。在这里，我有许多的人要感谢，没有他们，我不可能有现在的成长和成就。

首先，我要感谢我的毕设指导老师——赵云波教授，没有他在每个时间点上对我的督导，我将不能这么圆满地完成毕设。赵老师给我提供了实验室完成毕业设计的环境，也是赵老师细心地对我的文章提出修改意见并耐心地指导我文章架构的组织，在每次答辩的时候会给我们鼓励。

其次，我要感谢实验室的李灏学长。因为我的课题关于深度学习，对于计算机的配置要求较高，是李灏学长提供给我他的计算机，也是李灏学长帮助我配置好了深度学习的运行环境。如果没有他，我将无法顺利完成每个模型的训练与测试，在此表示衷心的感谢。

我也要感谢我的室友们，他们是陪伴我四年成长的人，在我面临学习上、生活上压力的时候给予我很多帮助。

最后感谢的是我的父母，他们孜孜不倦地工作，只为培养我成为一个优秀的大学生。他们常常给我打电话，关心我的学习和生活。我深深地能感受到父母对我的爱是无穷的、无私的。现在，我特别想成为一个足够有智慧、有担当、有勇气的人，这样就可以不再让父母担心了。

四年的磨砺，成长了曾经幼稚的我，我知道现在的我还不够优秀与成熟，但我相信，通过不断努力，我一定可以做的更好。再次感谢在成长路上帮助我的所有人。