硕士学位论文

基于深度学习的地铁非法过闸机行为检测研究与实现

Research and Implementation of Subway Illegal Turnstile Crossing Behavior Detection Based on Deep Learning

一级学科　控制科学与工程

二级学科　模式识别与智能系统

作者姓名　万思远

指导导师　谢斌 副教授

2022 年 5 月

中图分类号 TP391.4　　学校代码 10533

UDC 004.8　　学位类别 学术学位

硕士学位论文

基于深度学习的地铁非法过闸机行为检测研究与实现

Research and Implementation of Subway Illegal Turnstile Crossing Behavior Detection Based on Deep Learning

作 者 姓 名：　万思远

一 级 学 科：　控制科学与工程

二 级 学 科：　模式识别与智能系统

研 究 方 向：　深度学习

二级培养单位：　自动化学院

指 导 教 师：　谢斌 副教授

论文答辩日期 2022.5.25　　答辩委员会主席

中 南 大 学

2022 年 5 月

学位论文原创性声明

本人郑重声明，所呈交的学位论文是本人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除了论文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得中南大学或其他教育机构的学位或证书而使用过的材料。与我共同工作的同志对本研究所作的贡献均已在论文中作了明确的说明。

申请学位论文与资料若有不实之处，本人承担一切相关责任。

学位论文作者签名：万思远　　日期：2022 年 5 月 27 日

学位论文版权使用授权书

本学位论文作者和指导教师完全了解中南大学有关保留、使用学位论文的规定：即学校有权保留并向国家有关部门或机构送交学位论文的复印件和电子版，允许本学位论文被查阅和借阅。本人授权中南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和公开传播，可以采用复印、缩印或其它手段保存和汇编学位论文。本人同意按《中国优秀博硕士学位论文全文数据库出版章程》规定享受相关权益。本人保证：毕业后以学位论文内容发表的论文作者单位注明中南大学；学位论文电子文档的内容和纸质学位论文的内容相一致。

延缓公开论文延缓到期后适用本授权书，涉密论文在解密后适用本授权书。

本学位论文属于：(请在以下相应方框内打“√”)

☑公开

□延缓公开，延缓期限（___年__月__日至___年__月__日）

学位论文作者签名：万思远　　指导教师签名

日期：2022 年 5 月 27 日　　日期：2022 年 5 月 27 日

基于深度学习的地铁非法过闸机行为检测研究与实现

摘要：本文提出了一种基于深度学习的非法过闸机行为检测算法，并设计了一个基于原始图像和对应骨架图像的非法过闸机行为识别网络，目的是检测地铁闸机口的人员是否存在非法过闸机行为，并对发生非法过闸机行为的人员进行相应的行为告警。本文的具体研究工作可以分为以下几个部分：

（1）本文使用 YOLOv5 和 AlphaPose 在地铁闸机口构建了一个基于原始图像的非法过闸机行为识别数据集和基于骨架信息的非法过闸机行为识别数据集，为研究非法过闸机行为识别打下了基础，有利于后续更深入的研究。

（2）本文对非法过闸机行为识别数据集进行了基于原图和骨架图的行为识别研究，设计了一个基于原始图像和骨架图像的 OSFANet 网络，实验结果表明，本文设计的 OSFANet 网络有更大的优势，相较于单种数据的预测网络和中期融合网络有更好的分类效果，并做了消融实验验证 OSFANet 的有效性。

（3）本文设计并实现了一种基于视频事件的非法过闸机行为检测算法，对闸机口待检测区域内使用区域入侵算法进行行人检测，对检测到的行人进行对应的骨架提取，使用 OSFANe 进行非法过闸机行为识别，使用基于滑动窗口的方法对多帧图片中的人员进行跟踪并判断该人员是否存在非法过闸机行为，对正常过闸机的人员使用蓝色框框出即可，对发生非法行为的人员用红色框框出并保存为图片上传至对应的报警文件夹。使用了 10 个视频中的共 564 个非法过闸机行为事件做测试，其中平均正检率为 92.29%，平均召回率为 89.18%。

图 37 幅，表 13 个，参考文献 88 篇

关键词：行为识别；目标检测；骨架提取；注意力机制；信息融合
分类号：TP391.4

Research and Implementation of Subway Illegal turnstile Crossing Behavior Detection Based on Deep Learning

Abstract: This thesis proposes an illegal turnstile crossing behavior detection algorithm based on deep learning, and designs an illegal turnstile crossing behavior recognition network based on the original image and the corresponding skeleton image, which is utilized to detect whether the passengers have illegal turnstile crossing behavior, and give the corresponding behavior alarm to the passengers. This thesis can be divided into the following parts:

(1) In this thesis, YOLOv5 and AlphaPose are used to build an illegal turnstile crossing behavior recognition dataset based on the original image and an illegal turnstile crossing behavior recognition dataset based on skeleton information at the subway turnstile, which lays a foundation for the study of illegal turnstile crossing behavior recognition and is conducive to further research.

(2) This thesis studies the behavior recognition of illegal turnstile crossing behavior recognition dataset based on the original image and skeleton image, and designs an OSFANet based on the original image and skeleton image. The experimental results show that the OSFANet designed in this paper has greater advantages, and has better classification effect than the prediction network by single data and medium-term fusion network . Ablation experiments are carried out to verify the effectiveness of OSFANet.

(3) This thesis designs and implements an illegal turnstile crossing behavior detection algorithm based on video events. The area intrusion algorithm is used to detect pedestrians in the area to be detected at the turnstile, the corresponding skeleton of the detected pedestrians is extracted, the illegal turnstile crossing behavior is identified by osfane, and the method based on sliding window is used to track the person in multiple frames and judge whether the person has illegal turnstile crossing behavior, For the personnel who pass the turnstile normally, use

the blue box, and for the personnel who commit illegal acts, use the red box and save it as a picture and upload it to the corresponding alarm folder. A total of 564 illegal turnstile crossing events in 10 videos were used for the test, of which the average positive detection rate was 92.29% and the average recall rate was 89.18%.

Keywords: Behavior identification; Target detection; Skeleton extraction; Attention mechanism; Information fusion

Classification： TP391.4

目 录

第 1 章 绪论 1

1.1 研究背景及意义 1

1.2 国内外研究现状 2

1.2.1 目标检测研究现状 3

1.2.2 行为识别研究现状 6

1.3 论文的主要研究内容和贡献 9

1.4 论文的组织结构 10

第 2 章 非法过闸机行为识别数据集 12

2.1 研究动机 12

2.2 基于 YOLOv5 的行人检测 12

2.2.1 目标检测技术介绍 12

2.2.2 YOLOv5 目标检测算法研究 14

2.3 基于 AlphaPose 的骨架提取 15

2.3.1 骨架提取技术介绍 16

2.3.2 AlphaPose 骨架提取算法研究 16

2.4 非法过闸机行为识别数据集的建立 18

2.5 评价指标 19

2.6 本章小结 20

第 3 章 基于原图和骨架图的非法过闸机行为识别研究 21

3.1 研究动机 21

3.2 基于不同网络的非法过闸机行为识别 21

3.3 添加注意力模块的非法过闸机行为识别 24

3.3.1 不同注意力模块介绍 25

3.3.2 添加注意力模块的 ResNet50 网络设计 28

3.3.3 添加不同注意力模块实验结果对比分析 30

3.5 本章小结 32

第 4 章 结合原图和骨架图的非法过闸机行为识别研究 33

4.1 研究动机 33
4.2 基于信息融合的非法过闸机行为识别 33
4.2.1 前期融合算法设计 33
4.2.2 中期融合算法设计 34
4.2.3 后期融合算法设计 34
4.2.4 不同信息融合方式实验结果对比 35
4.3 基于 OSFANet 网络的行为识别 36
4.3.1 OSFANet 网络框架设计 36
4.3.2 OSFANet 主干网络设计 38
4.3.3 超分支 OSFA 模块设计 38
4.3.4 实验结果分析 41
4.3.5 消融实验分析 42
4.4 本章小结 44
第 5 章 基于视频行为事件的非法过闸机行为检测系统设计 45
5.1 研究动机 45
5.2 基于视频的非法过闸机行为检测算法的设计 45
5.2.1 基于区域入侵的行人检测 46
5.2.2 基于 OSFANet 网络的行为识别 49
5.2.3 基于滑动窗口的行人跟踪设计 50
5.3 基于视频事件的非法过闸机行为检测算法实现 52
5.4 本章小结 54
第 6 章 总结与展望 55
6.1 总结 55
6.2 展望 56
参考文献 57
攻读学位期间主要研究成果 64
致 谢 65

缩写索引

SIFT	Scale-invariant feature transform
HOG	Histogram of Oriented Gradient
DPM	Deformable Part Model
SVM	Support Vector machines
NMS	Non Maximum Suppression
YOLO	You Only Look Once
SSD	Single Shot Multibox Detector
STIP	Space-Time Interest Points
HOF	Histograms of Oriented Optical Flow
MBH	Motion Boundary Histograms
CNN	Convolutional Neural Network
Ori	Original image
Skeleton	Skeleton image
SENet	Squeeze-and-Excitation Networks
BAM	Bottleneck Attention Module
CBAM	Convolutional Block Attention Module
BTNK	Bottleneck
MHI	Motion History Image
DMM	Depth Motion Map
OSFANet	Ori And Skeleton Fusion With Attention Net
OSF	Ori And Skeleton Fusion
MFA	Mutiscale Fusion With Attention

第 1 章 绪论

1.1 研究背景及意义

科学技术是第一生产力，近年来人工智能技术在计算机领域被广泛关注。而计算机视觉技术在人工智能领域占据着重要地位。目前计算机视觉领域比较主流的研究方向有很多，目标检测，行为识别，骨架提取，目标跟踪等都是计算机视觉领域比较热门的技术。早期的机器视觉技术局限性很大，近年来，随着深度学习的热度不断提高，计算机视觉领域开始使用深度学习的方法，自从深度学习应用于计算机视觉领域，计算机视觉技术取得了重大突破。为计算机视觉带来了很大的贡献。虽然深度学习里的卷积神经网络在静态图片上的识别有着很大的优势，但是针对人体行为检测以及目标实时跟踪等基于动态视领域的研究还是比较有局限性，目前阶段还是很难到达可应用的地步。

行为检测不同于静态的图像分类问题，它是一个基于视频里的图像序列进行处理的问题。准确的讲是属于视频分类的情况。行为检测关注于视频中的目标对象随着时间的变化在空间中的动作变化。目的是为了能够理解人类通过动作而表现出来的具体行为。因此，行为检测的研究在人工智能领域上有着非常好的潜在价值和研究意义。由于行为检测算法是一种基于三维空间对数据进行处理的技术，技术会更加的复杂。但是维度增加的同时也增加了更多的方法可作为研究的切入点。近年来国内外有很多针对行人检测技术的相关研究，但是由于实时性和精确度还未能达到理想的情况，目前该技术的广泛应用还是需要一定的时间。从理论层面来说，行为检测其实就是用深度学习的方法在视频序列图片上进行图像分类，然后结合时间序列的信息综合进行的判断。虽然深度学习中的卷积神经网络在处理静态的图片时有着与众不同的优势，对静态的图片可以实现一个非常好的分类效果，但是针对动态的视频的分类效果和静态图像相比，还是有很大的差距。因为行为检测不仅要让计算机理解静态的图像上面有什么，还需要让计算机理解该序列图片具体在做什么，所以相对于静态图片的分类更具有挑战性。除非深度学习在针对时间序列这一维度上能取得重大突破，能实现一种很好的分类模型，针对视频序列的分类也能有一个很好的效果。在现实生活中，行为检测技术在各个领域还是非常具有应用价值的。比如在视频监控[1-3]领域，现实生活中有很多非法行为，像翻越闸机，偷窃纵火，打架斗殴等都属于非法行为，很多时候执法人员不一定能及时赶到，这时候如果能通过摄像头对这些非法行为进行识别，并及时进行报警，可以很大程度上降低损失。又比如在智能家居系统中，通过行为检测技术对用户的某些行为进行检测，在用户需要帮助的时候迅速地进行某些智能

服务，让用户在生活体验上有很大的提高。在医院里，可以通过行为检测技术对病人进行实时监控，一旦发现病人有危险行为立即启动报警，从而预防病人出现意外的风险。因此，无论是从研究的角度还是从实际生活层面来说行为检测技术都是非常重要的，如果行为检测技术能够在精确度和实时性上取得重大突破，对整个人工智能领域的发展将会迎来重大的推动。

近年来，随着我国经济发展水平逐渐提高，我国的城市化建设覆盖率也越来越广，大量的人口不断涌入城市，造成城市的公共交通非常的拥挤，而一线城市的交通更是堵塞严重。随着地铁在大城市的普及程度也越来越高，在很大程度上分摊了地面交通的人流量[4]。很多时候衡量一个城市的发展水平的重要方式之一就是看当地的地铁发展水平。由此可见，地铁在整个城市交通领域有着至关重要的地位。而地铁的安全建设是地铁安全[5]的保障，因此在地铁安置摄像头实施监控是地铁建设安防工作必不可少的一环。而乘客在进入地铁和走出地铁时都要经过闸机，同时闸机也是用于检验乘客是否买票的重要设备。而有一些乘客出于各种原因，在没有买票的情况下，通过一些非法的行为来通过闸机，这是非常不安全的。因此，地铁闸机口的安检建设尤为重要，如果地铁安检运行的不好，不仅给地铁公司带来不好的声誉同时也带来了很多的经济损失[6]。由于人力有限，很多时候并不能发现这些非法行为，此时我们就需要视频监控技术来对地铁的非法行为进行检测。因此，针对地铁场景的非法过闸机行为检测技术尤为重要，该技术不仅可以提高地铁里的乘客素质，还能加强地铁系统的安防监控能力。

本文的研究属于与海信公司合作的横向课题，主要是研究基于深度学习的地铁非法过闸机行为检测，首先，对地铁内闸机的监控视频流切分为图像序列，并定位每个闸机口的待检测区域；然后引入目标检测算法和骨架提取算法，构建每个闸机口人员的行为信息；再结合原图信息和骨架信息构建识别模型对人员的行为信息进行识别；最后设计相应的跟踪算法与逻辑规则生成相应的报警信息。本研究能够对非法过闸机行为进行很好的识别和跟踪，大大提升了地铁闸机口处人员通行的自动化监控与告警。

1.2 国内外研究现状

近年来，地铁闸机的检测技术主要还是在闸机口两侧安装红外传感器，通过红外线探测是否有行人通过。这种检测方式非常局限，只限于正常过闸机的情况，如果乘客使用非法行为跨越闸机的情况就无法感应，因此在地铁里针对非法过闸机行为检测技术尤为重要。随着计算机视觉技术和人工智能的飞速发展，对图像和视频的处理能力的愈发增强，以监控视频为基础的行为分析也取得了突飞猛进的发展，因此，通过视频分析技术判断过非法过闸机的行为的也具有

一定的可行性。随着以卷积神经网络为代表的深度学习的快速发展。基于单张图片的方法进行行人行为分析已广泛应用于现实生活中，给人们的生产生活带来了极大的便捷。像目标检测技术，骨架提取技术，行为识别技术等取得了许多有价值的成果[7]。其中，目标检测算法已经广泛用于生活场景图像检测与识别；行为识别技术也同样开始广泛应用于人的行为和姿态的分析识别。但是，当前的目标检测和行为识别技术很多是基于单帧图片的，无法从视频中的一系列连续动作判断人员的行为，而本文的非法过闸机的行为检测技术正是在这些技术的基础上进行研究的。

1.2.1 目标检测研究现状

目标检测技术概括来说就是一种对场景中的目标进行定位和识别的技术。我们实际生活中很多场景都会用到目标检测技术。比如在城市智能交通领域对马路上的行人和车辆进行检测，在密集的人群中对人头进行检测从而实现人数的清点，在安检时对对危险物品进行检测。从某种意义上来说，目标检测就是我们先选定一些感兴趣的事物，然后设计一个目标检测算法对图像中的事物进行识别，并找到该事物在图像中坐标位置。实质上来讲就是一种多任务学习，其中一个任务就是用具体坐标来描述目标在图中的区域位置，另一个任务是用一种分类算法对区域当中的事物进行识别分类。

在二十世纪九十年代，由于计算机的运算能力有限，人们只能通过提取图像中的一些特征通过对比相似度来作为检测。David Lowe[8]在 1999 年提出了一种尺度不变特征转换(SIFT)的算法，该算法可以用来检测图像中的局部性特征。尺度不变特征转换就是在尺度空间中找到极值点，并且提取其尺度，位置，旋转不变量等特征与数据库当中的特征做对比，通过特征向量当中的欧氏距离找到相似的特征，根据这种方法来实现检测。Viola-Jones[9]于 2001 年提出了一种用于人脸检测的算法，该算法结合了 Haar-like 特征、积分图像、Adaboost 和级联分类器等多种技术。第一步是通过在输入图像上滑动一个窗口来搜索类似 Haar 的特征，并使用积分图像进行计算。然后它使用经过训练的 Adaboost 来找到每个 Haar 特征的分类器并将它们级联。Viola Jones 算法是一种准确且功能强大的检测器。在 2005 年，Dalal 和 Triggs 提出了定向梯度直方图（HOG）[10]的方法，用于提取图像特征以进行目标检测。这是对其他检测器[11-14]的改进。HOG 提取梯度及其边缘方向以创建特征表。图像被划分为网格，然后使用特征表为网格中的每个单元创建直方图。为感兴趣的区域生成 HOG 特征并输入线性 SVM 分类器进行检测。该检测器被提出用于行人检测，并且它可以被训练来检测各种类别。Felzenszwalb 等人在 2008 年介绍了可变形零件模型(DPM)[15]。并且是 2009 年 Pascal VOC 挑

战赛的获胜者。它使用物体的单个“部分”进行检测，并获得了比 HOG 更高的准确度。它遵循分而治之的哲学；在推理期间单独检测对象的一部分，并将它们的可能排列标记为检测。例如，人体可以被认为是头部、手臂、腿和躯干等部分的集合。将分配一个模型来捕获整个图像中的一个部分，并对所有这些部分重复该过程。然后，模型会删除这些部分组合的不太可能的配置以产生检测。基于 DPM 的模型[16,17]是深度学习时代之前最成功的算法之一。

从 2012 年开始，随着 GPU 的运算能力显著提升，目标检测领域开始广泛地使用深度学习技术。起初人们使用两阶段法做目标检测，先选择目标区域再使用卷积提取特征。Girshick[18]等人的 R-CNN 算法，是第一个提出如何使用卷积神经网络来极大地提高检测性能的。该算法对输入的图像使用选择性搜索提取大约 2000 个候选区域对象，使用 AlexNet[19]作为检测器的主干网络对这些候选区域进行特征提取，然后将特征向量传递给经过训练的支持向量机(SVM)以获得置信度分数。用非极大值抑制（NMS）来拒绝 IoU 低于最高值的区域 。一旦确定了类别，该算法使用经过训练的边界框回归器预测其边界框，该回归器预测四个参数，即框的中心坐标及其宽度和高度。R-CNN 在目标检测领域迎来了新浪潮，但它在时间和空间上成本高。它有复杂的训练过程，即使在共享一些计算的情况下，也需要数天时间才能在小型数据集上进行训练。He 等人提出一种 SPP-Net[20]算法，在该算法被提出前，输入的图片大小都是固定的。但是他们意识到只有 CNN 的全连接部分需要固定输入。因此该算法用空间金字塔池（SPP）[21]来处理任意大小或纵横比的图像。SPP-Net 只是在区域提议模块之前移动了 CNN 的卷积层并添加了一个池化层，从而使网络可以输入任意大小/纵横比的图片，并减少了计算量。该算法先用选择性搜索算法用于生成候选窗口。将图像输入 ZFNet[22]的卷积层从而获得特征图。然后将候选窗口映射到特征图，随后通过金字塔池化层的空间箱将其转换为固定长度。该向量被传递到全连接层，最终传递给 SVM 分类器以预测类别和得分。与 R-CNN 类似，SPP-Net 有一个后处理层，通过边界框回归来改善定位。它也使用相同的多阶段训练过程，只是微调只在全连接层上进行。SPP-Net 比 R-CNN 模型快得多，具有相当的准确性。它可以处理任何形状/纵横比的图像，从而避免由于输入翘曲而导致的对象变形。然而，由于其架构类似于 R-CNN，它也具有 R-CNN 的缺点，例如多阶段训练、计算量大和训练时间长。为了解决 R-CNN 和 SPP-net 上述的问题，Girshick 提出 Fast R-CNN[23]算法，该算法通过创建一个单一的端到端可训练系统来解决这个问题。该网络将图像及其目标区域作为输入。图像通过一组卷积层获得特征图，并将目标区域映射到它。用 RoI 池化层替换了 SPP-net 中池化层的金字塔结构。该层连接到 2 个全连接层，然后分支为 N+1 类 SoftMax 层和一个边界框回归层，并且该层也有

一个全连接层。该模型还将边界框回归器的损失函数从 L2 更改为平滑 L1 以提高性能，并引入了多任务损失来更好地训练网络。该算法还使用了现有最先进的预训练模型的修改版本，如[24-26]作为主干网络。该网络通过随机梯度下降（SGD）和 2 个图像的小批量一步训练。这有助于网络更快地收敛，因为反向传播在来自两个图像的 RoI 之间共享计算。Fast R-CNN 在提高了准确率的同时速度也比 R-CNN 快 146 倍，该算法的优点在于简化了训练过程，去除了金字塔池化并提出了新的损失函数。没有区域提议网络的目标检测器报告了接近实时的速度，并且具有相当的准确性。尽管 Fast R-CNN 更接近实时目标检测，但它每张图片的候区域生成需要 2 秒，并不够快速。Ren[27]等人提出 Faster R-CNN 算法，该算法本质上是使用 RPN 作为候选区域的 Fast R-CNN。由于两个执行非常不同任务的模型之间存在共享层，Faster R-CNN 的训练更加复杂。Faster R-CNN 修复了候选区域生成慢速的瓶颈，并将推理时间减少了一个数量级（以每秒 5 帧的速度近乎实时地运行）。并且检测精度比之前的 Fast R-CNN 的检测技术提高了 3%以上。

在实际生活中，检测技术要达到应用的层面，实时性尤为重要，两阶段检测器的优点在于精度高，但是由于两阶段检测需要先选定候选区域再进行分类，实时性上还有待提升。为了弥补速度上的缺陷，人们开始研究单阶段检测器。Redmon[28]提出 YOLO 算法，是第一个把深度学习应用在单阶段目标检测的。YOLO 将检测视为回归问题，直接将图像像素预测为对象及其边界框属性。YOLO 的灵感来自用于图像分类的 GoogLeNet 模型[29]，该模型使用较小卷积网络的级联模块[30]。YOLO 在准确性和速度上都以巨大的优势超越了其当代的单级实时模型。然而，它也有明显的缺点。小型或聚集对象的定位精度以及每个单元格对象数量的限制是其主要缺点。SSD[31]是第一个与 Faster R-CNN 等当代两级检测器的精度相匹配的单级检测器，同时保持实时速度。SSD 基于 VGG-16[32]构建，具有额外的辅助结构以提高性能。在训练期间，SSD 将每个 ground truth 框与具有最佳 Jaccard 重叠的默认框匹配，并相应地训练网络，类似于 Multibox[33]。它还使用硬负挖掘和大量数据增强。与 DPM[34]类似，它利用定位和置信度损失的加权和来训练模型。通过执行非最大抑制获得最终输出。尽管 SSD 比 YOLO 和 Faster R-CNN 等最先进的网络明显更快、更准确，但它在检测小物体方面存在困难。随后，Redmon 等人[35]对 YOLO 进行了改进，提出了 YOLOv2 网络。该网络对速度和准确性进行了折中，用 DarkNet-19[36]替换了 GoogLeNet 的主干架构。并且它结合了许多令人印象深刻的技术，如批量标准化[37]以提高收敛性、分类和检测系统的联合训练以增加检测类别、移除完全连接的层以提高速度。紧接着提出 YOLOv3 网络[38]，该算法用更大的 Darknet-53[36]网络替换了特征提取器网络。他们还结合了各种技术，如数据增强、多尺度训练、批量标准化等。分

类器层中的 Softmax 被逻辑分类器取代。尽管 YOLOv3 比 YOLOv2 更快，但它与其前身相比缺乏突破性的变化。在 2021 年，YOLOv4[39]利用数据增强技术、正则化方法、类标签平滑、CIoU-loss[40]、自对抗训练、余弦退火调度器[41]和其他技巧来改进训练。它还使用遗传算法搜索超参数。它有一个 ImageNet 预训练的 CSPNetDarknet-53 主干、SPP[37]和 PAN 路径聚合块[42]以及 YOLOv3 作为检测头。大多数现有的检测算法都需要多个 GPU 来训练模型，但是 YOLOv4 可以很容易地在单个 GPU 上进行训练。YOLOv5 仅在其前身（即 YOLOv4）之后一个月由一家名为 Ultralytics[43]的公司发布，对现有的 YOLO 检测器进行了多项重大改进。YOLOv5 在速度和准确性方面都优于两阶段目标检测算法。该算法是最好的目标检测算法之一，被广泛地应用在实际生活中。

综上所述，目标检测技术经历了从传统图像处理技术到深度学习网络的转变过程。而传统图像处理技术非常有局限性，着重关注于图像本身的特征，对图像中光照不均，复杂背景等干扰情况下检测结果不稳定。而深度学习中的卷积神经网络正好能解决这些问题。在深度学习阶段。目标检测又经历了从两阶段到一阶段的转变。如上述所说，两阶段采用定位加分类的方法，检测速度和单阶段比还存在差距，随着计算机技术的不断发展，目标检测的性能会取得更好的效果。而本文的非法过闸机行为检测不存在小目标检测的问题，一般都能检到，在保证精度的情况下对实时性要求更高，考虑到工程应用价值本文选择单阶段目标检测算法 YOLOv5 做为本文的行人检测算法。

1.2.2 行为识别研究现状

人体的行为信息是社会互动的重要组成部分，基于计算机视觉和机器学习的人体行为识别已成为智能监控、人机交互、智能家居、虚拟现实、医疗诊断等诸多不同领域的研究热点。行为识别的过程通常被认为是特征提取、特征描述和分类器的结合[46-54]。特征提取的方法在识别过程中至关重要，因为特征提取的好坏直接影响最终识别结果的准确性。根据特征提取的方法，行为识别方法可以分为传统的人工特征方法和深度学习方法。人类行为识别早期采用传统的人工特征方法[44-48]。在该方法中，专家通过手工设计特征的方式对视频的底层特征进行提取，然后将低层特征预处理并抽象为高维特征，最后编码为特征向量[55-57]。向量作为分类器的输入，通过训练过程最终得到一个好的分类器。深度学习方法则可以自动提取人类行为视频中的行为特征向量，形成高维行为特征，以便更好地进行行为识别，其中通过数据标签和预测值建立损失函数，深度学习模型的超参数不断通过反向传播算法进行调整，直到获得良好的识别结果。深度学习方法可以适应复杂场景中常见的光照和遮挡问题[58,59]。

在行为识别的早期发展时期，普遍采用传统的人工特征提取方法。根据提取特征的不同，基于人工特征的方法可以进一步分为基于时空兴趣点（STIP）的方法[49,50,54]、基于模板匹配的方法[56]基于轨迹的方法[64]、基于语义空间的方法[65]和基于深度序列的方法[66,67]。针对基于时空兴趣点的方法，在 2003 年，Laptev[49]等人通过将时域参数添加到 Harris 2D 兴趣点将空间兴趣点扩展到时空域，然后提出 3D Harris STIP，并使用由兴趣点组成的点集来表示动作。该方法实际上检测到运动方向相反的点，因此检测效果在周期性和旋转运动中效果不佳。此外，Harris 角点检测算法的特征点相对稀疏，在复杂的真实场景中容易导致提取信息不足，识别效果不佳[60]。为了解决这个问题，Dollar 等人[50]提出了一种 Cuboid STIP 检测算子，对空间和时间域的信息采用不同的处理方法，空间域信息采用高斯核函数，时间信息提取采用一维 Gabor 滤波器。对周期性运动的物体或具有明显运动特征的物体表现出良好的检测效果。Scovanner 等人[54]提出的 3D 尺度不变特征变换（SIFT）检测算子，可以准确地获得时空特征，对图像缩放、旋转甚至仿射变换具有鲁棒性。Willems 等人[61][54]提出了一种基于 3D Hessian 矩阵的时空兴趣点检测方法，已被证明具有高效计算。针对基于模板匹配的方法，Blank 等人[62][57]在 2005 年提出了一种 MEI 的扩展方法，该方法使用 3D 人体在时间和空间上的形状来描述动作，从而增加了视点变化的鲁棒性。但是，为了进行分类识别，需要将 3D 人体轮廓替换为 2D 平面，这是通过计算每个点到达边界的平均时间来实现的。Yilmaz 和 Shah[55]提出根据时空体积（STV）的不同属性确定动作。物体的外部轮廓沿时间轴堆叠在一起，形成一个 STV。Klaser 等人[56]在 2008 年将静态图像的梯度直方图（HOG）特征扩展到时空维度，并提出 3D HOG 特征来表示动作。该方法不需要特征点检测，但需要密集采样视频来提取 3D HOG 特征。后来 Sahoo 等人[63]提出了一种基于 MHI 的兴趣点细化算法（IPRA），将方向梯度直方图和光流直方图技术从空间扩展到时空域，以节省时间信息。针对基于轨迹的方法，在 2012 年，Wang 等人[57]提出了密集轨迹（DT）算法，DT 方法通过网格划分在多个尺度上对图像的特征点进行密集采样，然后在去除一些无法跟踪的缺少变化的特征点后形成密集采样的关键点。下一步就是沿着时间轴通过整个视频跟踪视频序列中每一帧密集采样的关键点，形成密集的轨迹。最后，将灰度图像直方图（HOG），光流直方图（HOF）和运动边界直方图（MBH）三个描述符组合起来形成特征描述符，用于表示人体信息。Wang 等人[68]在 DT 算法的基础上提出了一种改进的密集轨迹（IDT）算法。与 DT 算法相比，IDT 消除了背景光流（估计相机运动），因此特征更侧重于描述人体运动。Yi 等人[69]提出了一种基于显着性检测和低秩矩阵恢复的显着前景轨迹提取方法（SFRE）。密集轨迹分为显着轨迹和非显着轨迹。Haiam 等人[70]提出了一种改进的基于轨迹

的人体动作识别方法。通过选择 Dollar 提出的长方体检测器[50]，这种改进的方法克服了稀疏 STIP 检测器只能检测少量稳定兴趣点的缺点。基于运动轨迹的方法的优点是这些方法全面描述了整个运动过程，对快速不规则运动具有很强的鲁棒性。缺点是密集采样导致特征维度高，计算量大，并且还包含不相关的背景信息。针对基于语义空间的方法，Zhu[65]提出了三种基于语义主题模型的识别方法，即隐马尔可夫主题模型（HMTM）、协同聚类主题模型（CCTM）和主题隐马尔可夫模型（THMM），以改善基于语义的方法的不足。HMTM 弥补了 PLSAM 和 LDAM 等语义主题模型只关注语义信息的缺点。它是语义顶层模型（STM）和时间关系隐马尔可夫模型（HMM）的组合，同时关注语义信息和时间关系。这种方法的优点是可以通过语义分析来弥合计算机和人脑之间的语义鸿沟，而缺点是特征的高维表示限制了计算速度。针对基于深度序列的方法，陈等人[66]提出了用于人体动作识别的深度运动图（DMM），该方法通过将每个深度帧的前视图、侧视图和顶视图投影到三个正交笛卡尔平面上来获得三个 2D 投影图像。将两个相邻深度帧之间的运动能量差叠加起来形成 DMM 特征向量。杨等人[67]提出了 DMM-HOG 方法，类似于 Chen 等人的 DMMs 方法[65]。它们之间的区别在于 DMM-HOG 使用定向梯度直方图 HOG 描述符对 DMM 进行编码，并探索在识别任务中可以达到最佳性能的帧数。深度序列法的优点是受环境因素影响较小，鲁棒性高，缺点是容易忽略帧间的动作信息，容易导致数据冗余。

随着深度学习的发展和大数据时代的临近，深度学习已经在图像领域得到应用。随着取得巨大成功，深度学习在基于视频的行为识别中的应用逐渐成为一种发展趋势。Ji 等人[71]在 2010 年提出了一种基于 3D CNN 的行为识别方法，Tran 等人[72]在 3D CNN 的基础上提出了卷积 3D（C3D）网络。与 3D CNN 不同的是，这种方法使用完整的视频帧作为输入，而不是分段的视频，因此更容易扩展到大型数据集。随后，Tran 等人[73]提出了一种新的 Res3D 网络方法，将残差网络[74]与 C3D 网络相结合。网络模型的大小是 C3D 的一半，但运行速度是 C3D 的两倍。C3D 网络既可以从视频的连续帧中提取 3D 卷积特征，也可以提取时间维度和空间维度的特征信息。由于可以同时处理多帧图像，因此运行速度加快，但计算成本大，对硬件要求较高。因此一些研究人员提出在预训练的 2D 网络的基础上制作 3D 卷积网络。例如，Qiu 等人[75]提出了一种伪 3D 残差网络（P3D ResNet），将 1*3*3 空间卷积层和 3*1*1 时间卷积层结合起来，以取代标准的 3*3*3 卷积，其大小减少了模型的体积，简化了 3D 卷积。2014 年，Simonyan[76]提出了一种基于双流网络结构的行为识别方法。研究中提出的双流网络结构分为两个独立的识别部分，即时间流 CNN 和空间流 CNN。时间流 CNN 使用多帧堆叠的光流图像作为输入来处理时间维度信息，而空间流 CNN 使用单帧 RGB 图像作为输入来

处理空间维度信息。将两个分支的输出向量融合，得到最终的分类结果。该方法由于加入了时间流 CNN，提高了视频行为识别的准确率。Zhao 等人[77]提出了一种改进的双流结构，其中选择长短期记忆网络而不是 CNN 作为空间流，以从空间流中提取空间和时间信息。与其他方法相比，该方法获得了更高的准确率，但由于其复杂度高且参数较多，训练过程耗时较长。Li 等人[78]提出了一种基于骨架数据的分层聚合共现特征学习方法，该方法将输入分成两个流，一个流使用 CNN 对每个骨架关节点学习点级特征，然后通过分层特征学习网络将点级特征逐渐聚合成全局共现特征。另一个流生成时差序列，对时差序列执行相同的点级特征学习。然后它聚合成一个全局的共现特征，最后将两个流融合在一起。Tang 等人[79]提出了一种基于骨架的深度渐进强化学习方法用于人类行为识别，该研究提出了帧蒸馏网络，使用深度强化学习来提取视频中信息量最大的帧；并且提出了一种基于图的 CNN（GCNN），通过构建图 G(x,W)表示人体骨骼的拓扑结构，它更好地表达了关节之间的依赖关系和空间相关性。在 2017 年，Rahmani 等人[80]提出了一种基于深度和骨骼数据的端到端深度人类行为识别模型。该方法通过双线性紧凑池有效地结合深度数据和骨架数据，可以解决视图变化导致的深度图像不一致的问题。2020 年，Annalisa 等人[53]充分利用 Kinect 传感器设计一种结合骨架和 RGB 数据流的多模态行为识别方法，Pushpajit 等人[81]提出了一种基于 CNN 的多模态数据组合方法。该方法将 RGB 数据构造的 MHI、深度图像数据构造的 DMM（包括前视图、侧视图和顶视图）以及由骨架序列构造为 5 个分支的骨架图像分别输入预处理后的 CNN 进行训练，最后通过加权乘积法进行决策层融合。这种方法可以充分利用 RGB-D 传感器获取的数据，即使数据集出现错误，仍然可以进行识别工作。基于特征融合的行为识别算法可以综合多种不同特征的优势，充分利用多尺度的行为信息。缺点是不同特征之间的融合方式对结果影响较大，需要寻找合适的融合方式。

综上所述，行为识别经历了从传统人工特征方法到深度学习方法的转变过程。数据也更加多样化，从最初的单一数据到现在的多信息融合。早期的传统人工特征方法需要依靠专家设计并提取特征，存在更多的局限性，后来引入深度学习方法，可以通过卷积神经网络自动提取特征，还可以适应复杂场景中常见的光照和遮挡问题。随着计算机技术的不断发展，行为识别技术将会得到更广泛的关注。

1.3 论文的主要研究内容和贡献

本文主要是研究基于深度学习的地铁非法过闸机行为检测，首先，对地铁内闸机的监控视频流进行切分为图像序列，并定位每个闸机口的图像范围；其次，引入目标检测算法和骨架检测算法，构建每个闸机口人员的行为信息；再者，结

合设计的识别模型对人员的行为信息进行识别；最后，设计相应的跟踪算法与逻辑规则生成相应的报警信息。本研究能够适应更加广泛的场景，能够对非法过闸机行为进行很好的识别和跟踪，大大提升了地铁闸机口处人员通行的自动化监控与告警。本文的主要研究内容和贡献包括：

（1）本研究使用 YOLOv5 目标检测技术构建了一个地铁闸机口的非法过闸机行为识别的数据集，并且使用 AlphaPose 骨架提取技术构建了基于骨架的非法过闸机行为识别数据集，为研究非法过闸机行为识别打下了基础，有利于后续更深入的研究。

（2）本文对非法过闸机行为识别数据集进行了基于原图和骨架图的分类研究，设计了一个基于原始图像和骨架图像的 OSFANet 网络，实验结果表明，本文设计的 OSFANet 网络有更大的优势，相较于单种数据的预测网络和中期融合网络有更好的分类效果，并做了消融实验验证 OSFANet 的有效性。

（3）本文设计了一个基于深度学习的非法过闸机行为检测算法，针对地铁场景下的闸机口非法过闸机行为有着很好的检测效果，并且该算法能够适应更加广泛的场景，能够对非法过闸机行为进行很好的识别和跟踪，大大提升了地铁闸机口处人员通行的自动化监控与告警。

1.4 论文的组织结构

本文包含六个章节，文章的组织结构如图 1-1 所示，各个章节内容介绍如下：

第一章首先介绍了基于深度学习的地铁非法过闸机行为检测的研究背景及意义。然后介绍了目标检测算法的研究现状和行为识别算法的研究现状。最后阐述了本文的主要研究内容及贡献和论文的组织结构。

第二章主要研究对地铁闸机口的非法过闸机行为识别数据集的建立，通过 YOLOv5 的目标检测算法对地铁闸机口的行人进行检测，对图像中闸机口检测框内的行人进行保存，形成基于原始图像的非法过闸机行为识别数据集。然后使用 AlphaPose 骨架提取技术对检测框的图像进行骨架提取，构建了基于骨架的非法过闸机行为识别数据集。

第三章主要研究基于原图和骨架图的非法过闸机行为识别，使用多种分类网络对两种数据集进行了非法过闸机行为识别的实验并进行对比分析，然后对最好的分类网络分别添加不同注意力机制进行实验对比，验证注意力机制对提升分类网络的有效性。

第四章主要研究结合原始图像和骨架图像的非法过闸机行为识别，使用信息融合技术进行非法过闸机行为识别，比较前期融合，中期融合和后期融合的效果，随后提出了一种 OSFANet 网络，经过实验对比和消融实验，OSFANet 网络对非

法过闸机行为识别的效果更好。

第五章阐述了基于深度学习的地铁非法过闸机行为检测算法的研究。对闸机口的行人进行基于区域入侵的行人检测和行为识别后采用基于滑动窗口的跟踪技术综合多帧结果判断此人是否有非法过闸机行为。对有非法过闸机行为的人员进行监控和报警。

第六章对本文的主要研究内容进行相应的总结，然后分析并讨论提出的算法尚存在的不足点，对下一步地铁非法过闸机行为检测算法的优化和研究进行了展望。

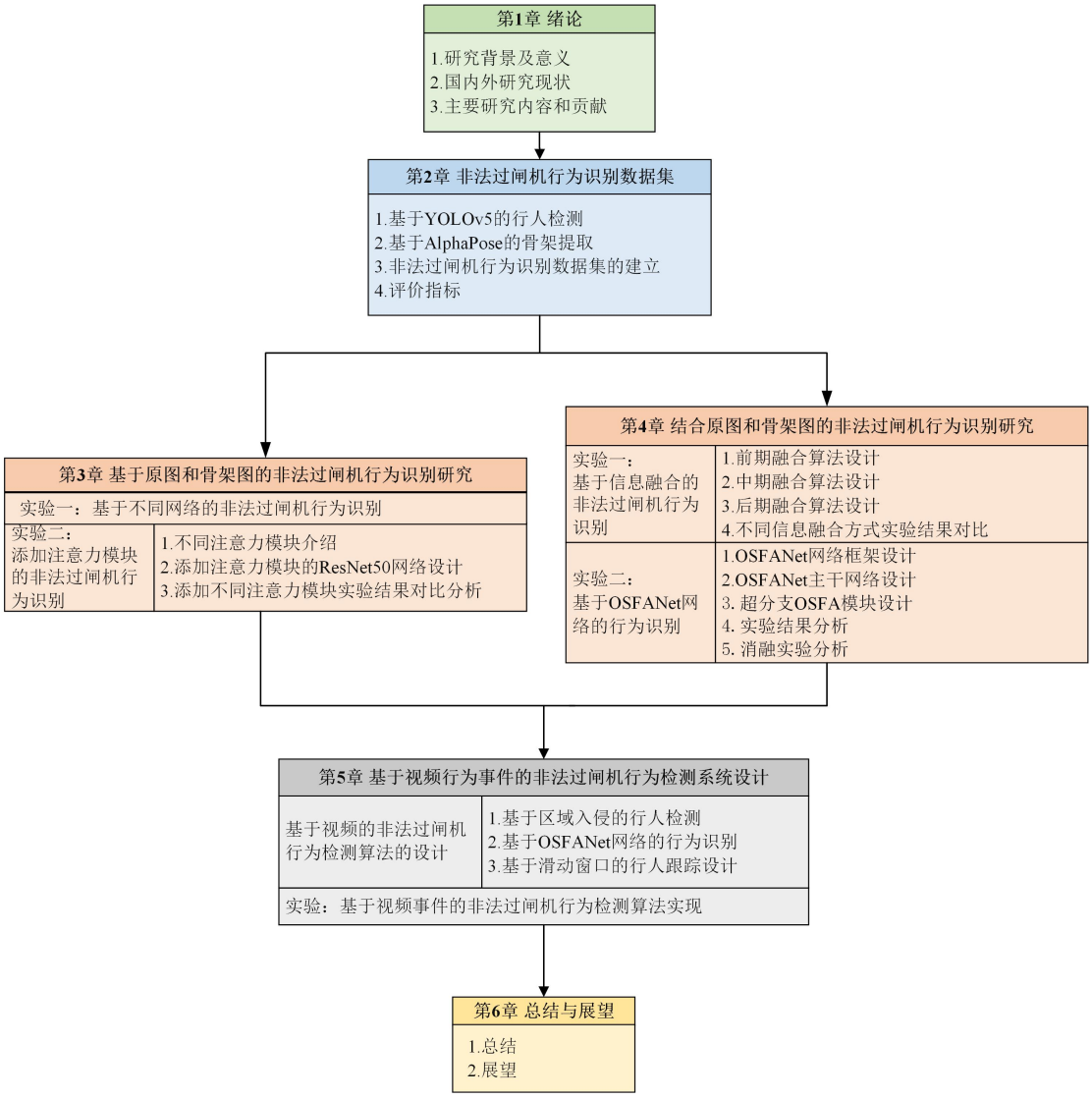


图 1-1 论文组织结构

第 2 章 非法过闸机行为识别数据集

2.1 研究动机

针对监控视频中的非法过闸机行为检测，本文的思路是先对视频流里的单帧图片进行非法过闸机行为识别，然后对视频序列图片设计相应的跟踪算法和逻辑规则实现非法过闸机行为检测。因此我们需要建立一个非法过闸机行为识别数据集来实现单帧图片的非法过闸机行为识别。

2.2 基于 YOLOv5 的行人检测

2.2.1 目标检测技术介绍

目标检测技术被广泛地应用在智慧交通，工业检测，人脸识别等领域。和图像识别不同，目标检测需要对图像中的事物进行定位并且进行分类。目标检测的任务量要比图像分类多，因此检测的难度要比分类大。随着科学技术的发展，目标检测技术已经从传统的图像处理技术发展到应用深度学习技术的重大转变，总体来说，目标检测的类型如图 2-1 所示。

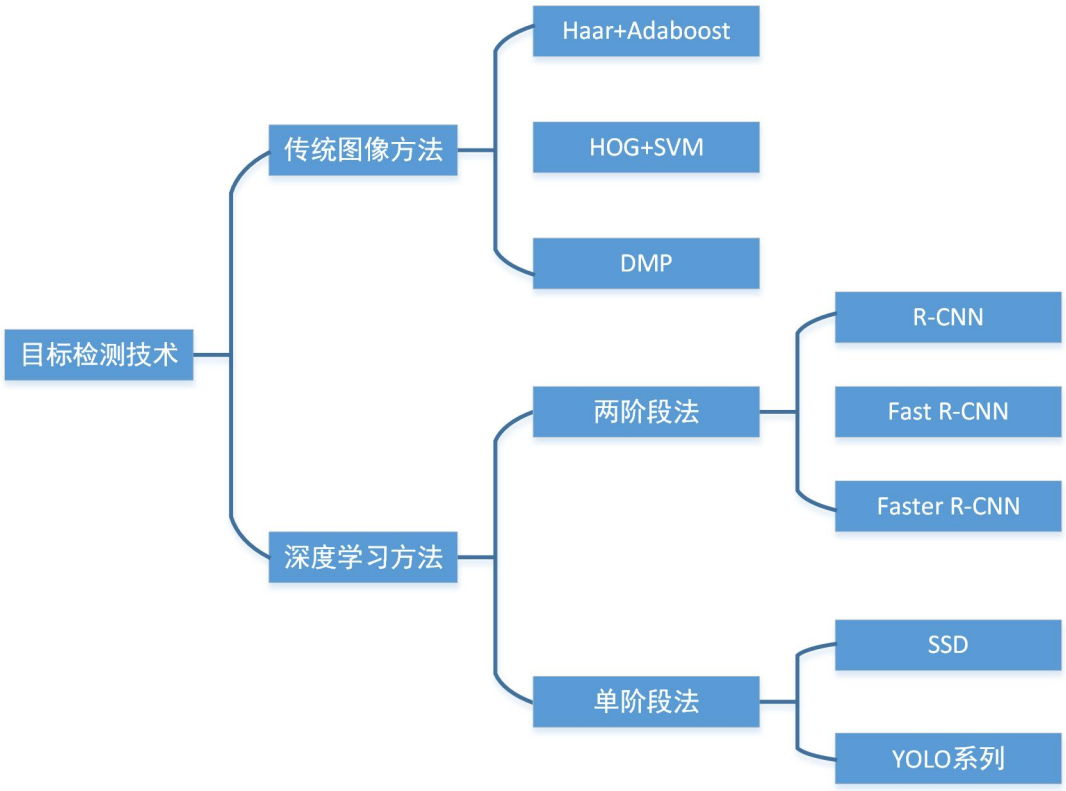


图 2-1 目标检测类型

早期的目标检测技术都是传统的图像处理技术，人们只能通过提取图像中的一些特征通过对比相似度来作为检测。无论是 Haar-like 特征加 Adaboost 和级联

分类器的方法还是 HOG 提取梯度及其边缘方向以创建特征表加 SVM 分类器进行检测又或者是基于 DPM 的模型的方法，都需要人工提取特征，然后采用滑动窗口的方法对图片进行特征搜索，这种方法泛化能力低，并且对图像中光照不均，复杂背景等干扰情况下检测结果不稳定。而深度学习中的卷积神经网络正好能解决这些问题。基于深度学习的目标检测方法主要分为两大类，一类是先对图像生成候选区域，然后对通过 CNN 对它们进行分类加定位，这种我们称为两阶段目标检测。如图 2-2 所示。

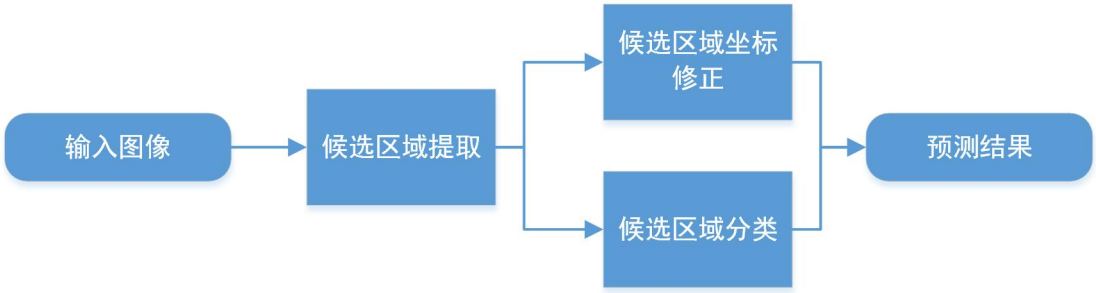


图 2-2 两阶段目标检测原理

另一种没有生成候选区域，而是结合回归的思想，用 CNN 直接对图像进行分类和定位，这种称为单阶段目标检测，如图 2-3 所示。两阶段目标检测精度较好，但是训练比较复杂，速度上也不是很快。而单阶段目标检测在速度上有很大的优势，并且准确率也不错，适合应用在实际生活中。本研究采用单阶段目标检测对闸机口的行人进行检测。

图 2-3 单阶段目标检测原理

本文研究的基于深度学习的地铁非法过闸机行为检测技术第一步就是需要对地铁闸机口的乘客进行行人检测，行人检测技术是实现非法过闸机行为检测的基础。经过调研目标检测的发展现状我们已经知道两阶段目标检测精度虽然高，但是耗时较长，单阶段目标检测可以在保持准确性的情况下速度上有很大的提升。目前单阶段目标检测比较主流的算法是 SSD 和 YOLO 系列，而 YOLO 系列在工程性和易用性上比其它目标检测算法要好，近年来 YOLOv5 网络更是得到了广泛地应用，具有很高的实用性。由于项目对实时性的要求比较高，因此，本文选用 YOLOv5 目标检测算法进行行人检测的研究。

2.2.2 YOLOv5 目标检测算法研究

YOLO 系列目标检测的框架，通常可以分为如下几个部分：输入端、骨干网络、Neck 网络和输出端。目前 YOLOv5 算法共有 5 个版本，分别是：YOLOv5s、YOLOv5m、YOLOv5l、YOLOv5x 和 YOLOv5n。本节以 YOLOv5s 为例进行说明，其他版本的结构原理与之类似。YOLOv5s 目标检测算法的整体网络结构如图 2-4 所示。YOLOv5[38]的基础组件由 CBL，Res unit，CSP1_X，CSP2_X，Focus 和 SPP 组成，YOLOv5 的结构可分为输入端，骨干网络。Neck 网络和 Head 输出端四大组成部分。

（1）输入端表示输入的图片，该网络的输入图像大小为 608*608，该阶段通常包含一个图像预处理阶段，即将输入图像进行归一化等操作，使图像与网络的输入大小保持一致。在训练过程中，YOLOv5 使用 Mosaic 数据增强操作提升模型的训练速度和网络的精度。

（2）骨干网络通常是一些性能优异的分类网络，该模块用来提取一些通用的特征表示。YOLOv5 中不仅使用了 CSPDarknet53 结构，而且使用了 Focus 结构作为基准网络。

（3）Neck 网络通常位于基准网络和头网络的中间位置，利用它可以进一步提升特征的多样性及鲁棒性。虽然 YOLOv5 同样用到了 SPP 模块、FPN+PAN 模块，但是实现的细节有些不同。

（4）Head 输出端用来完成目标检测结果的输出。针对不同的检测算法，输出端的分支个数不尽相同，通常包含一个分类分支和一个回归分支。YOLOv5 利用 GIOU_Loss 来代替 Smooth L1 Loss 函数，从而进一步提升算法的检测精度。

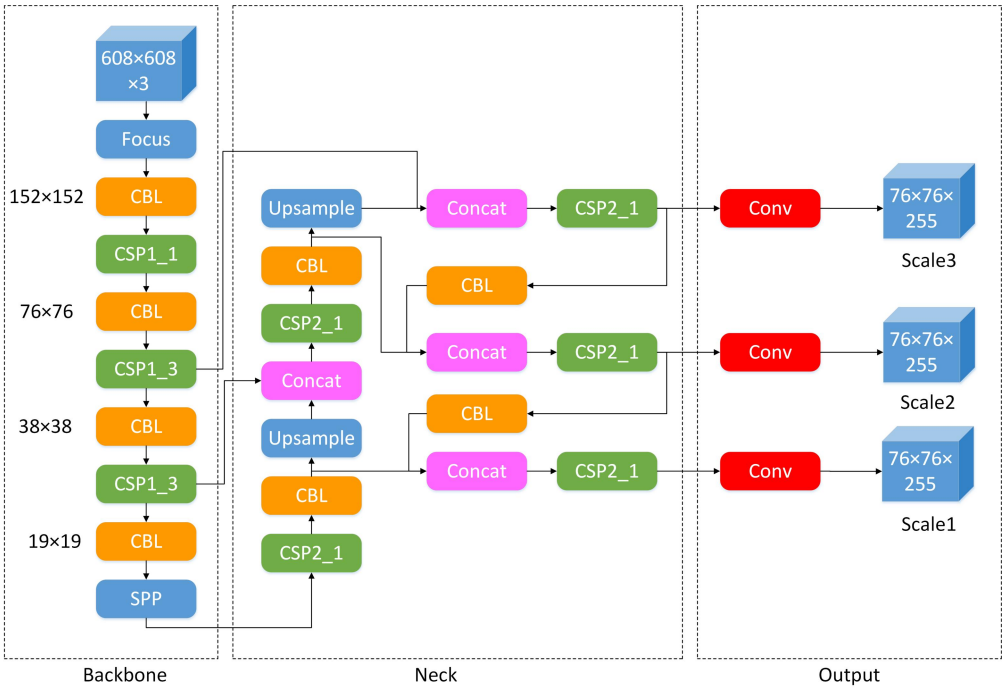


图 2-4 YOLOv5 的整体网络结构

针对海信公司提供的视频数据，在经过切分为视频序列图片后，本文并没有使用 YOLOv5 目标检测算法直接进行行人检测，而是使用了一种基于区域入侵的目标检测算法（本文 5.2.1 会做详细介绍），首先对闸机口区域进行定位，在图片中人工选取了四个坐标点，形成了一个四边形区域，而地铁闸机正处于该区域中。而图片中的行人有些处于区域内，有些处于区域外，本文只对区域内的人员进行行人检测。为了获得一个针对地铁闸机口场景的行人检测模型，本文首先使用经过开源的 COCO 数据集进行行人检测训练的 YOLOv5 预训练模型，从而获得比较好的初始值并且缩短了训练时间，然后使用手工标注的 1000 张地铁闸机口的行人数据进行再训练，从而获得一个针对地铁闸机行人检测的模型，用这个 YOLOv5 模型进行地铁闸机口的行人检测。YOLOv5 算法会计算出图像中所有人员检测框的四个顶点坐标，我们对检测框中对角的两个点进行区域判断，如果这两点同时满足在四边形区域内的情况，则对该人员进行检测，如果不满足条件则不做检测处理。行人检测的效果如图 2-5 所示。

由图 2-5 可知，本文之所以选用左上角偏大的四边形，是根据摄像头的角度和行人过闸机前的路径综合决定的，针对不同的闸机口，设定的四边形区域也会有所不同，像图中的情况，摄像头位置有所偏移，闸机口靠墙，行人有从左侧进入的视角更多，因此对左上角划分更大的位置，从而更好地检测闸机口的人员，也为后续进行非法过闸机行为检测提供更足的判断时间。

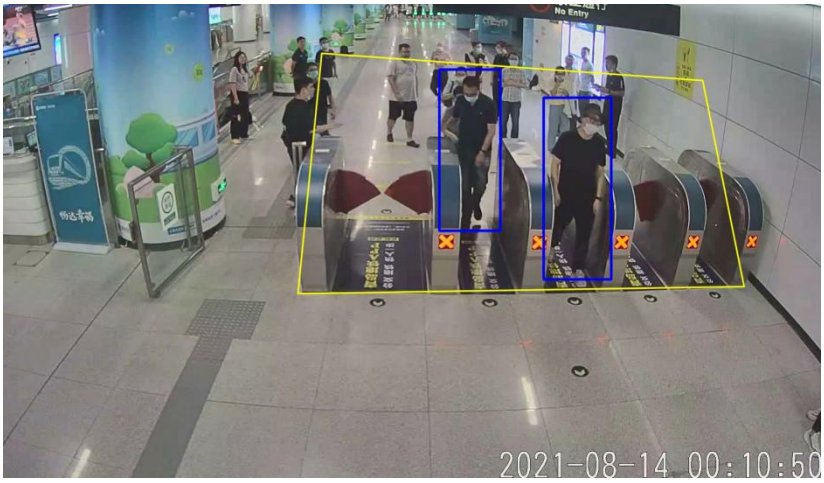


图 2-5 基于 YOLOv5 的行人检测

用 YOLOv5 目标检测算法对区域内进行行人检测后，本文会对检测框中的人员图片进行保存，用这些图片构成基于原始图像的非法过闸机行为识别数据集。

2.3 基于 AlphaPose 的骨架提取

为了实现非法过闸机行为检测算法，本文从多个方向进行非法过闸机行为识别的研究。而骨架数据就是一个很好的行为识别研究方向，因此针对地铁场景下

闸机口行人过闸机时 YOLOv5 算法所检测到的行人，使用 AlphaPose 算法进行骨架提取，从骨架信息来对非法过闸机行为进行识别。

2.3.1 骨架提取技术介绍

人体骨架提取技术又称为人体骨骼关键点检测技术，是一种通过算法来检测人体头部的关键点（双眼，双耳，鼻子），手部的关键点（双肩，双手肘，双手）和腿部关键点（双臀，双膝盖，双脚）等的主要关键点，通过这些关键点的连接来描述人体骨架信息。该技术被广泛应用在体育运动，智能安防，视频监控，智慧医疗，行为识别等领域。早期的骨架提取技术大多是根据几何先验知识，通过手工设计特征的方法来进行表示，然后通过模板匹配的方法进行人体关键点检测。该方法存在许多缺点，其中最主要的问题在于模板的选取，需要找到一个合适的模板能够表示人体的整体结构，并且其中的关键点的连接能够展示各肢体之间的位置关系。然而很多手工特征选取的模板并不能很好地展示人体骨架的位置信息，并却泛化性能较差，在特定的数据集上效果还行，迁移到别的数据集就不行。而基于深度学习的骨架提取技术很好地弥补了手工特征的缺陷，该方法不仅可以能够在图片中自动学习并提取骨架特征，而且在检测速度上和准确性上都优于传统的手工特征方法。基于深度学习的骨架提取技术大体分为两类，一种是自上而下[82]的方法，另外一种是自下而上[83]的方法。自上而下的人体骨架提取技术采用的是一种分治思想，先对图像上的人进行目标检测，然后对单个的人进行骨架提取。流程如图 2-6 所示。

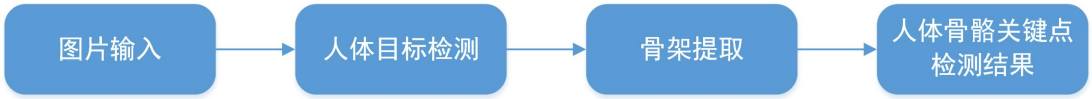


图 2-6 自上而下的人体骨架提取流程

自下而上的方法则是先把图像中的关键点检测出来，然后对关键点进行聚类。流程如下图 2-7 所示。

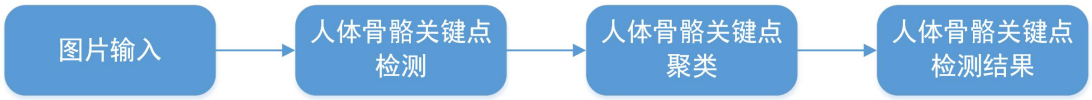


图 2-7 自下而上的人体骨架提取流程

本文考虑到地铁闸机口的行人需要先进行行人检测，因此选用自上而下的方法。

2.3.2 AlphaPose 骨架提取算法研究

AlphaPose 网络是一种自上而下的骨架提取网络，该网络在 2017 年由腾讯依

图和上海交大一起提出，入选了当年的 CVPR。该网络主要是解决了工程上的两大难题，一种是检测框的质量不是很好，另一种是检测框会出现重复的情况，因此提出了一种新的骨架提取算法 AlphaPose，取得了优异的效果。该算法的整体模块结构如图 2-8 所示。

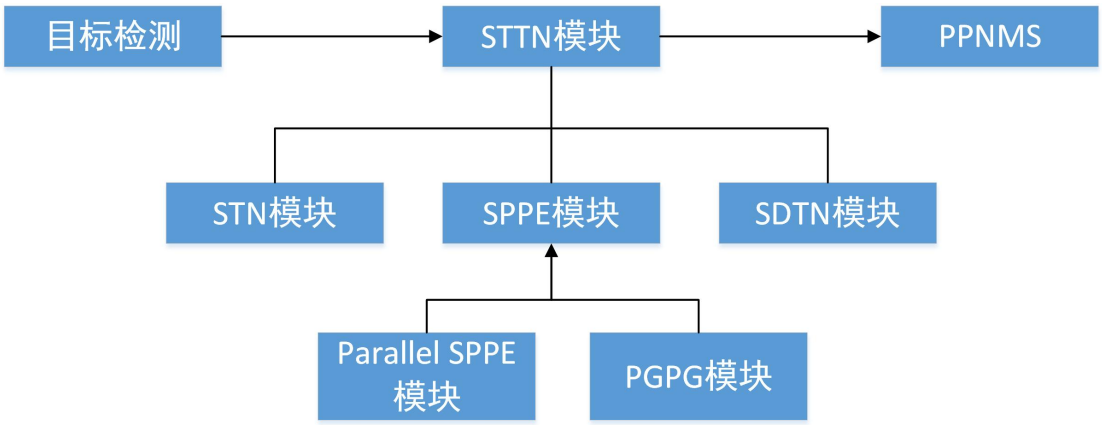


图 2-8 AlphaPose 整体模块结构

（1）STN 和 SDTN 分别为空间变换模块和空间反变换模块。STN 的作用是可以将输入的图片进行各种自适应地空间变换，从而提高分网络的准确性。通俗地讲，在训练阶段，STN 可以对输入的图片进行自动选择感兴趣的部位提取特征，从而实现对不同形变的数据进行空间变换。SDTN 和 STN 相反，属于空间反变换模块，在 AlphaPose 中的主要作用是将姿态估计后所得到的人体姿态重新映射到原始的图像坐标中，通过这个过程来训练 SPPE 姿态估计网络。STN 和 SDTN 这两种模块在 AlphaPose 算法中有着非常优异的性能，尤其是在自动选择感兴趣区域方面。因此使用 STN 来处理输入的人体区域框，在经过训练的 SPPE 后，通过 SDTN 将获得的姿势映射回原始的人体区域框中。

（2）Paralell SPPE 的主要作用是让 STN 提取出一个更好的人体框。该模块在测试阶段是不会使用的，主要是应用在训练过程中。在训练过程中 Paralell SPPE 和 SPPE 会同时处理 STN 模块输出的人体框，由于 STN 模块不能将姿态转换到人体框的中心位置，在 SPPE 姿态估计后的姿态返回人体框时的位置和真实标签的位置出现误差时，Paralell SPPE 会对这个误差进行反向传播到 STN，从而帮助 STN 能够聚焦在中心位置。换言之 Paralell SPPE 就相当于是训练过程中的正则化过程，从而减少出现局部最优的情况。因此 AlphaPose 算法通过 Paralell SPPE 模块提高 STN 模块将姿态估计后的人体姿态移动到检测框中间的能力，从而有效提升 SPPE 的姿态估计效果。

（3）PPNMS 是一种参数化姿态非极大值抑制方法。在进行人体检测定位的时候产生冗余的姿态检测或者是人体检测框是不可避免的情况，因此在输出的时

候进行姿态非极大值抑制来消除冗余的情况是很有必要的。但是普通的方法精度和效率并不是很好，而 AlphaPose 中的 PPNMS 定义了一种姿态距离，同时加入了带参数的 NMS,在迭代过程中固定两个参数变化另外两个参数来搜索最优解，如果收敛，在测试阶段这些参数将会固定，从而选择最优的姿态。

（4）PGPG 模块是一种数据增强的方法。对于先定位区域后姿态估计的这种双阶段姿态识别，适当的数据增强有助于 SPPE 姿态估计适应人体框定位不完美的结果。该模块会产生和人体检测结果一样的样本作为训练集，从而达到数据增强的效果，进一步提高算法的性能。

针对经过 YOLOv5 算法进行行人检测的检测框中图片，本文使用 AlphaPose 进行骨架提取，如图 2-9-a 所示为 YOLOv5 目标检测算法检测框内行人非法过闸机时的原始图片，对原始图片进行了骨架提取，获得经 AlphaPose 算法处理后的骨架图如图 2-9-b 所示。

(a)原图 (b)骨架图

图 2-9 原图和骨架图

2.4 非法过闸机行为识别数据集的建立

海信公司提供了 30 个地铁闸机口行人非法过闸机的视频，本文对海信公司提供的地铁闸机口监控视频先进行视频抽帧，本文按 1 秒 5 帧的速度进行抽帧后切分为图像序列，然后对图像中闸机口的监控区域进行定位，我们只检测定位区域的行人。我们把行人过闸机的行为定义为两类，一类是非法过闸机行为，另一类为正常过闸机行为。非法过闸机行为的定义为挡板关闭的情况下翻越闸机和跨越闸机，而正常过闸机行为的定义就是通过投币和刷卡等正常方式等挡板打开时走过闸机。首先使用 YOLOv5 对划定的四边形闸机区域内的行人进行检测，对检测后的行人进行裁剪并按照行人过闸机的行为定义存入非法行为和正常行为文件夹做为基于原始图像的非法过闸机行为识别数据集，再使用 AlphaPose 生成对应的骨架数据集。通过 YOLOv5 目标检测算法和 AlphaPose 骨架提取算法，本

研究获得了基于原始图像的非法过闸机行为识别数据集和基于骨架信息的非法过闸机行为识别数据集。其中骨架数据集作为原图数据集的辅助数据集，统称为非法过闸机行为识别数据集。在基于原图的数据集中，共获得 278479 张图片。其中非法过闸机行为图片为 14980 张，正常过闸机行为图片有 263499 张。在基于骨架的数据集中，共获得 267015 张图片。其中非法过闸机行为图片为 14547 张，正常过闸机行为图片有 252468 张。原图数据集和骨架数据集数量不一致的原因在于并不是每张原图都能生成对应的骨架图。因此骨架的数据集会比原图数据集的数量少一点，由于非法过闸机的数据集和正常过闸机数据集数据不平衡，本文将非法过闸机数据每帧都进行训练，正常过闸机数据每隔 8 帧选一张进行训练，我们通过这种挑选方式作为行为识别网络的训练数据集。数据集的制作流程图如图 2-10 所示。

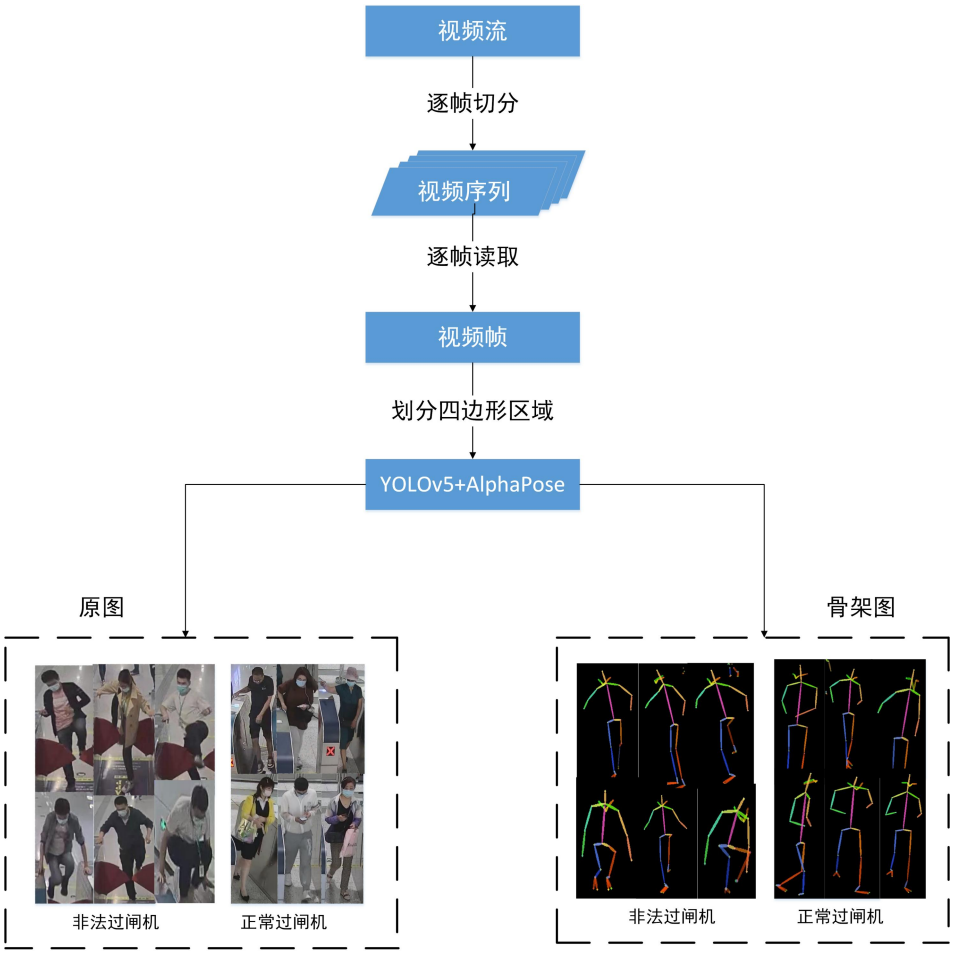


图 2-10 原图和骨架图数据集制作流程

2.5 评价指标

一般情况下，针对非法过闸机行为识别这种二分类问题，使用混淆矩阵来表

示。混淆矩阵针对结果主要有以下四种情况：预测为真，实际为真，这种又称为真阳性，用 True Positive 来表示，简称 TP；预测为真，实际为假，这种又称为假阳性，用 False Positive 来表示，简称 FP；预测为假，实际为真，这种又称为假阴性，用 False Negative 来表示，简称 FN；预测为假，实际为假，这种又称为真阴性，用 True Negative 来表示，简称 TN；如表 2-1 所示。

表 2-1 混淆矩阵分类结果

预测结果\实际结果	正例	负例
正例	TP	FP
负例	FN	TN

通过对上述 TP，FP，FN，TN 的数量进行统计，便可以计算出准确率、精确率、召回率和 F_1。计算公式如公式（2-1）、（2-2）、（2-3）和式（2-4）所示：

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2-1}$$

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

$$Recall = \frac{TP}{TP + FN} \tag{2-3}$$

$$F_1 = \frac{2 * Precision * Recall}{Precision + Recall} \tag{2-4}$$

准确率就是预测正确的数量占所有预测数量的百分比，在样本不平衡时尽量避免使用。精确率就是预测正例实际也为正例的数量占所有预测为正例的数量百分比，召回率就是预测为正例实际也是正例的数量占实际为正例的数量百分比，F_1 也叫 F1-Score，是精确率和召回率的调和平均值，是一种综合评价指标。

2.6 本章小结

本章首先介绍了研究非法过闸机行为识别数据集的研究动机，数据集的建立是研究非法过闸机行为检测的基础，因此研究非法过闸机行为识别数据集的建立非常必要。然后介绍了 YOLOv5 目标检测算法，该算法是数据集建立的第一步，也是后续研究非法过闸机行为检测算法的第一步，通过 YOLOv5 算法截取了闸机区域内的行人过闸机图片，形成基于原始图像的非法过闸机行为识别数据集。然后做了原图对应的骨架提取实验，在经过 AlphaPose 骨架提取算法后获得了对应的基于骨架的非法过闸机行为识别数据集。

第 3 章 基于原图和骨架图的非法过闸机行为识别研究

3.1 研究动机

从上一章的介绍中可知，通过目标检测技术和骨架提取技术，我们已经获得了两种地铁非法过闸机行为识别数据集。一种是基于原始图像的非法过闸机行为识别数据集，另一种是基于骨架图像的非法过闸机行为识别数据集。想要实现非法过闸机行为检测技术，在对闸机口的行人进行检测后，下一步就是对该乘客进行行为识别，这也是本文研究的核心技术。而行为识别技术非常广泛，国内外研究现状中已经具体介绍，如图 3-1 所示。

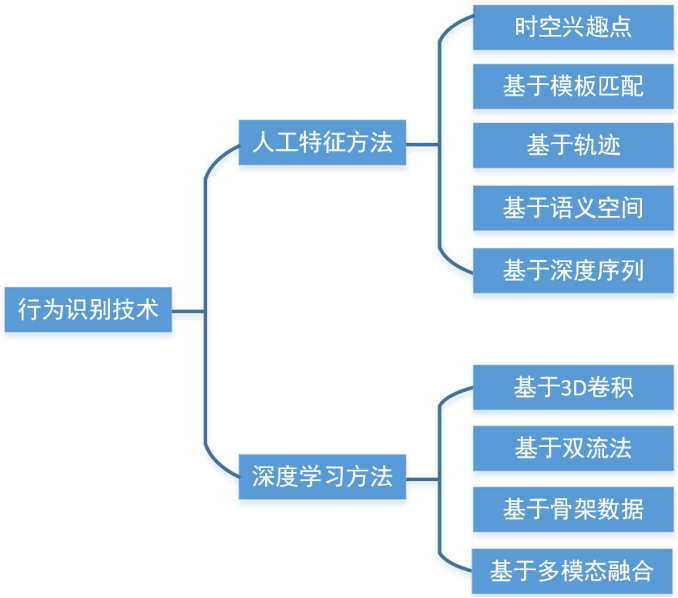


图 3-1 行为识别技术介绍

行为识别的好坏直接影响非法过闸机行为检测算法是否能正常运行。接下来本文会针对这两种数据集进行非法过闸机行为识别的研究。

3.2 基于不同网络的非法过闸机行为识别

卷积神经网络 (CNN) 是一种由多个阶段组成的人工神经网络，其中每个阶段本身就是一个小型神经网络,它被称为层。每一层都找到数据的一个给定特征，随着逐渐向后面的层移动，识别特征的复杂性会增加，直到我们可以识别完整的对象。一个 CNN 通常由 4 种类型的层组成：卷积层、池化层、ReLU 层和全连接层。

卷积层中卷积的主要目的是提取图像中的特征。卷积通过使用输入数据的小方块学习图像特征来保留像素之间的空间关系。这种小方块称为过滤器，通过在图像上滑动过滤器并计算点积形成的矩阵称为卷积特征。用不同滤波器进行卷积

所得到的效果也不同，我们可以更改滤波器矩阵的数值来执行边缘检测、锐化和模糊等操作。

池化层的作用主要是逐步减少输入空间的大小，在保留主要特征的情况下能够减少网络中的参数和计算的数量。池化的类型分最大值池化和平均值池化等。

ReLU 层可以将像素值为负数的值替换成零，使卷积神经网络引入非线性，可以解决线性模型不能解决的问题。

全连接层是传统的多层感知器，它在输出层使用 softmax 激活函数（也可以使用 SVM 等其他分类器），卷积层和池化层的输出代表输入图像的高级特征，全连接层的目的是利用这些特征根据训练数据集将输入图像分类为各种类别。

近年来，随着计算机的运算能力越来越强，越来越多优秀的卷积神经网络被提出，从最开始的 AlexNet，随后又出现了 VGGNet，再到后面的 GoogleNet 网络，网络越来越深，理论上讲，网络越深，对图像的特征提取效果会越好，但实际上随着网络的加深，在反向传播的梯度求导过程中很可能会出现梯度爆炸或者梯度消失的情况，从而导致训练过程的收敛效果不好，这时候网络太深反而不好，后来，ResNet 网络被提出，通过使用残差结构不但可以加深网络层数同时还能提升当前模型的性能，残差结构如图 3-2 所示。

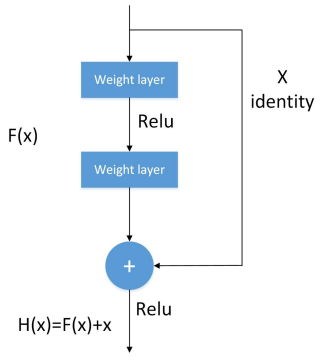


图 3-2 残差结构

本文首先尝试了使用原图和骨架图单独输入深度神经网络，来验证识别的准确性。本文首先分别使用了 VGG-16，GoogleNet，ResNet18，ResNet18，ResNet34，ResNet50 网络作为本文非法过闸机行为识别的训练网络。虽然还有更深的 ResNet101 和 ResNet152，考虑到层数太多导致运算量的增加是很耗时的，因此不做考虑。ResNet 网络中由许多残差块组成，这些残差块可以可以解决梯度爆炸和梯度消失的问题，从而加大网络的深度，提升网络的性能。其中 ResNet18 和 ResNet34 网络主要是由 Basic Block 组成，而 ResNet50 网络主要是由 BottleNeck Block 组成，如表 3-1 所示为 ResNet18，ResNet34，ResNet50 网络的组成结构。

表 3-1 ResNet18、ResNet34、ResNet50 网络的组成结构对比

模块名称	ResNet18	ResNet34	ResNet50
Conv1	Conv,7×7，/2	Conv,7×7，/2	Conv,7×7，/2
	3×3 max pooling,/2	3×3 max pooling,/2	3×3 max pooling,/2
Conv2-x	[Basic Block]×2	[Basic Block]×3	[BottleNeck Block]×2
Conv3-x	[Basic Block]×2	[Basic Block]×4	[BottleNeck Block]×4
Conv4-x	[Basic Block]×2	[Basic Block]×6	[BottleNeck Block]×6
Conv5-x	[Basic Block]×2	[Basic Block]×3	[BottleNeck Block]×3

由于非法过闸机的数据集和正常过闸机数据集数据不平衡，本文将非法过闸机数据每帧都进行训练，正常过闸机数据每隔 8 帧选一张进行训练，非法过闸机行为数据和正常过闸机数据集按照 1:2 的比例进行训练，训练周期为 50 个 epoch，训练的 batch size 为 128，学习率设为 0.0008，优化器为 Adam，上一次梯度更新的权重 momentum 设为 0.9，权重损失 weight decay 设为 1e-4。训练完成后，使用海信公司另外提供的 5 个非法过闸机的测试视频来进行测试。同样本文先对视频按 1 秒 5 帧的方式进行抽帧再使用 YOLOv5 和 AlphaPose 算法生成对应的原图和骨架图数据，对经过行人检测截取出来的图片和经骨架提取算法处理后的图片进行筛选，剔除不好定义的行人数据以及不正确的骨架数据，经过挑选和整理，最终获得 8284 张原始图片和对应的骨架图片，以此作为本文训练模型的测试集。该测试集中非法过闸机图片有 2445 张，正常过闸机有 5839 张。对训练后的模型用该测试集中的两种数据进行测试，模型的预测结果如表 3-2 和表 3-3 所示。表中 Ori 表示原始图像数据集（Original image），Skeleton 表示骨架数据集（Skeleton image）。

由表 3-2 和表 3-3 对比实验结果可知，ResNet50 对原图和骨架图的识别效果更好一点，而原始图像的识别效果比基于骨架的效果相对较好一点，其中主要原因是骨架提取本身就存在一定误差，并不准确，而原始图像是最真实的行为，因此原图的识别效果相对骨架来讲会更好一点。基于原始图像的 ResNet50 混淆矩阵分类结果如表 3-4 所示。

表 3-2 不同模型的预测原图结果对比

数据类型	模型	Acc(%)	Precision(%)	Recall(%)	F1-Score(%)
Ori	VGG-16	88.967	81.724	80.654	81.186
Ori	GoogleNet	89.039	81.743	80.941	81.340
Ori	ResNet18	88.955	81.664	80.695	81.177
Ori	ResNet34	89.353	82.495	81.145	81.814
Ori	ResNet50	**89.619**	**82.843**	**81.759**	**82.297**

表 3-3 不同模型的预测骨架图结果对比

数据类型	模型	Acc(%)	Precision(%)	Recall(%)	F1-Score(%)
Skeleton	VGG-16	87.844	79.516	79.223	79.369
Skeleton	GoogleNet	88.122	80.274	79.223	79.745
Skeleton	ResNet18	87.651	79.139	78.978	79.058
Skeleton	ResNet34	88.472	80.964	79.673	80.313
Skeleton	ResNet50	**88.725**	**81.362**	**80.164**	**80.758**

表 3-4 基于原始图像的 ResNet50 混淆矩阵分类结果

预测结果\实际结果	非法过闸机	正常过闸机
非法过闸机	1999	414
正常过闸机	446	5425

由混淆矩阵可知，由于负样本较多，所以准确率会偏高，而本文主要关注正样本非法过闸机行为的分类效果，因此，本文舍弃准确率作为评价指标，主要以精确率，召回率和 F1-Score 作为评价指标。

3.3 添加注意力模块的非法过闸机行为识别

随着深度学习地不断发展，受到人类注意力机制的启发，深度学习中同样可以使用注意力机制。近几年注意力机制在语音识别，机器视觉，信息融合等领域获得了广泛地应用。注意力机制可以对输入的大量信息进行处理，着重关注于关键信息区域，对无关紧要的冗余信息进行抑制，有效提取关键信息。在深度学习领域，注意力机制又可以根据计算区域分为硬注意力机制和软注意力机制。硬注意力机制的权重设置只有 0 和 1 两种，针对关注区域设置权重为 1，对不关注的区域设置为 0，由于硬注意力不可导，一般在训练的时候需要引入强化学习。软注意力和硬注意力不同，该方法对全局的特征权重设置范围在 0 到 1 之间，权重越大说明该特征越受关注，因此软注意力是可导的，添加软注意力的卷积网络可以反向传播。常见的软注意力计算方式由如下几种：

$$\text{加性模型：}\ att(x_i,q)=v^T\tanh(Wx_i+U_q) \tag{3-1}$$

$$\text{点积模型：}\ att(x_i,q)=x_i^Tq \tag{3-2}$$

$$\text{双线性模型：}\ att(x_i,q)=x_i^TWq \tag{3-3}$$

其中，W,U 和 v 表示可学习的网络参数，q 表示特定任务的相关参数信息，$att(x_i,q)$ 是 x_i 得到注意力计算的结果。Woo 等人[80]基于 ImageNet 数据集在 ResNet 网络上对比了不同注意力模型给网络的增益，因此本文针对非法过闸机行为识别引入注意力机制的研究。

3.3.1 不同注意力模块介绍

3.3.1.1 SENet 注意力模块

SENet[84]是 Hu 等人在 2017 年提出的一种即插即用的注意力模块，该模块可以在深层网络中学习特定的信息，一般在卷积模块的后面进行使用。SENet 模块的整体结构如图所示，该模块主要对经过卷积到的特征图进行以下三个操作：Squeeze(压缩)、Excitation(激励)、Scale。整体结构如图 3-3 所示。

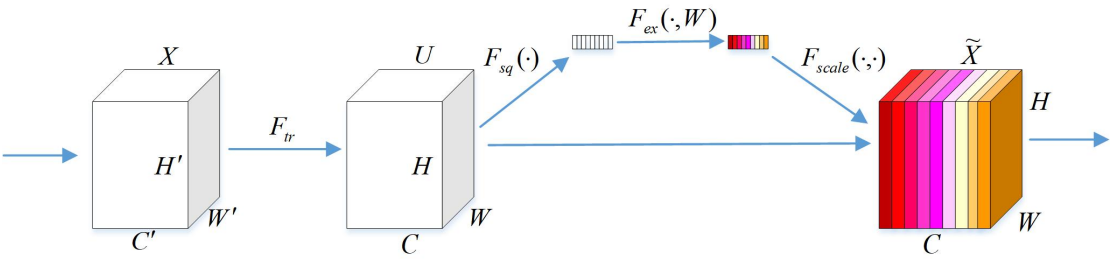


图 3-3 SENet 整体结构

在 Squeeze 操作中，主要是通过全局池化的方式把特征图卷积成 1×1×C 的向量，其中 C 表示通道数，通过这种方式我们就实现了空间上的压缩，其中没勇敢数代表着相应通道上的评价分数。操作如式(3-4)。

$$z_c = F_{sq}(u_c) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} u_c(i, j) \tag{3-4}$$

在 Excitation 操作中，为了在通道间建立相关性模型，使用两个全连接层并且在其中加入 ReLu 激活函数来对通道特征间进行非线性变换，经过学习来为各个特征通道生成权重。公式如下：

$$s = F_{ex}(Z, W) = \sigma(g(Z, W)) = \sigma(W_2 \delta(W_1 Z)) \tag{3-5}$$

在 Scale 操作中，就是将 Excitation 操作后的得到的通道权重与原特征图对应通道的二维矩阵相乘。即对原有的特称图进行通道方向上的加权。公式如下：

$$\tilde{x}_c = F_{scale}(u_c, s_c) = s_c u_c \tag{3-6}$$

3.3.1.2 BAM 注意力模块

BAM[85]注意力模块由通道注意力模块和空间注意力模块并联而成。整体结构如图 3-4 所示。

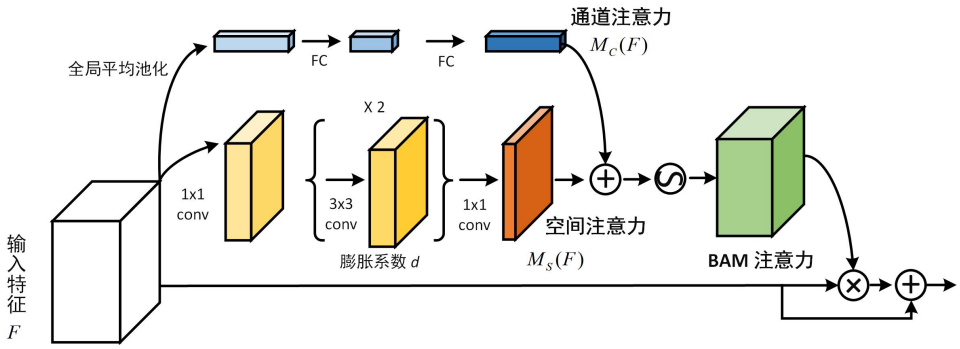


图 3-4 BAM 整体结构

BAM 注意力模块中的通道注意力可以将输入的特征 F 经过全局平均池化，输入到多层感知机（MLP）中，然后对多层感知机中输出的特征进行归一化操作，最后压缩成 C×1×1 的特征向量。公式如下：

$$M_C(F) = BN(MLP(AvgPool(F))) \tag{3-7}$$

BAM 注意力模块中的空间注意力是将输入特征 F 输入到 4 个卷积层中，第一个卷积是 1×1 的，通过这个卷积来压缩特征图的通道维数，后面两个卷积是 3×3 的卷积，主要目的是增加网络的有效感受野，这两层中的膨胀系数 d=4，最后一层卷积也是 1×1，主要是生成 1×H×W 的特征向量，最后也需要经过归一化操作。公式如下：

$$M_S(F) = BN(f_3^{1\times1}(f_2^{3\times3}(f_1^{3\times3}(f_0^{1\times1}(F))))) \tag{3-8}$$

由整体结构可知，设输入特征为 F，输出为 F'，M_C 表示通道注意力的特征映射，M_S 表示空间注意力的特征映射，用 σ 表示激活函数，则整体框架的数学表达式如下所示：

$$F' = F + F \otimes \sigma(M_C(F) + M_S(F)) \tag{3-9}$$

3.3.1.3 CBAM 注意力模块

CBAM[86]和 BAM 一样都使用了通道注意力模块和空间注意力模块，但是与 BAM 不同之处在于 BAM 的通道注意力和空间注意力模块是并联的，而 CBAM 则是串联的。CBAM 对输入特征先进行通道注意力模块处理然后进行空间注意力模块处理，整体结构如图 3-5 所示。

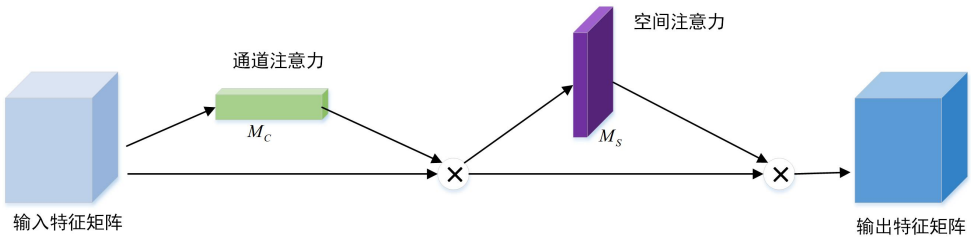


图 3-5 CBAM 整体结构

由 CBAM 的整体结构可以得知，设输入特征为 F，通道注意力的特征映射为 M_C，经过 CBAM 中通道注意力的输出特征为 F'，空间注意力特征映射为 M_S，经过空间注意力模块的输出为 F''，则 CBAM 的数学表达式为式(3-10)所示。

$$\begin{aligned} F' &= M_C(F) \\ F'' &= M_S(F') \end{aligned} \tag{3-10}$$

相比与 BAM，CBAM 模块在通道注意力模块和空间注意力模块上也有所优化。CBAM 的通道注意力模块为了能够提取更多的通道特征信息，在 MLP 前面增加了最大池化和平均池化层，同时使用两种池化层可以提取更精细的特征信息，使网络具有更好的表达能力。通道注意力模块如图 3-6 所示。

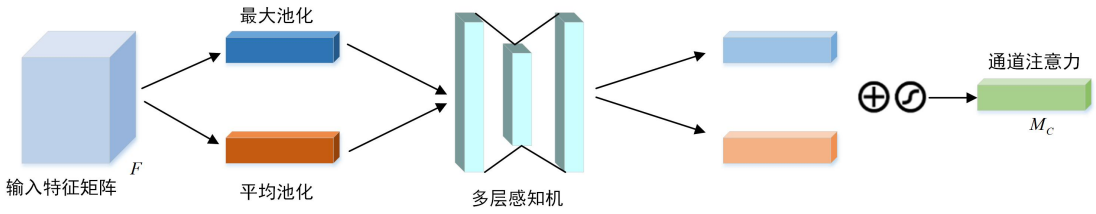


图 3-6 CBAM 的通道注意力模块

通道注意力模块的数学表达式如下：

$$M_C(F) = \sigma(MLP(AvgPool(F) + MLP(MaxPool(F)))) \tag{3-11}$$

由数学表达式和图可以得知，输入的特征 F 经过分别经过最大池化和平均池化后，输入一个共享的多层感知机，对多层感知机网络输出的两种特征进行叠加后输入激活函数中，最后获得 F'。

CBAM 的空间注意力模块也增加最大池化和平均池化，相比与 BAM 减少了 3 个卷积，整体结构如图 3-7 所示。

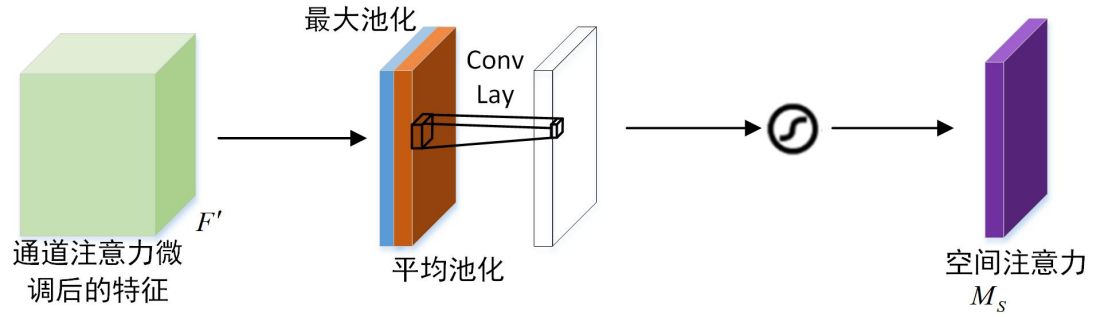


图 3-7 CBAM 的空间注意力模块

由图 3-8 可知，输入 F 先后通过最大池化和平均池化后，再输入一个 7×7 的卷积，最后输入到激活函数中。相应的数学表达式为：

$$M_S(F) = \sigma(f^{7\times7}([AvgPool(F); MaxPool(F)])) \tag{3-12}$$

3.3.2 添加注意力模块的 ResNet50 网络设计

经过对比实验我们已经知道 ResNet50 网络的识别效果较好，因此本文选择 ResNet50 网络作为添加注意力机制的网络，ResNet 网络的优点在于能够通过残差块来提升网络深度的同时避免梯度消失或者梯度爆炸的问题，从而更好地提取特征，从而提高网络的性能。如图 3-8 所示，ResNet50 网络的整体框架由 5 个阶段（stage1，stage2，stage3，stage4，stage5）组成，其中 stage1 相对简单一些，可以认为是对输入图像的预处理，（3,224,224）表示通道数为 3，宽度和高度为 224，Conv（Convolution 的缩写）是卷积操作，7×7 是卷积核的大小，64 是卷积核的数量，也就是经过卷积层后输出的通道数，/2 表示卷积核的步长为 2。BN 是 Batch Normalization 的缩写，表示批量标准化，就是常说的 BN 层。RELU 表示 ReLU 激活函数。MaxPool 表示最大值池化。后面 4 个 stage 的结构比较相似，都是由不同个数的 Bottleneck 组成，stage2 有 3 个 Bottleneck，stage3 有 4 个 Bottleneck，stage4 有 6 个 Bottleneck，stage5 有 3 个 Bottleneck。

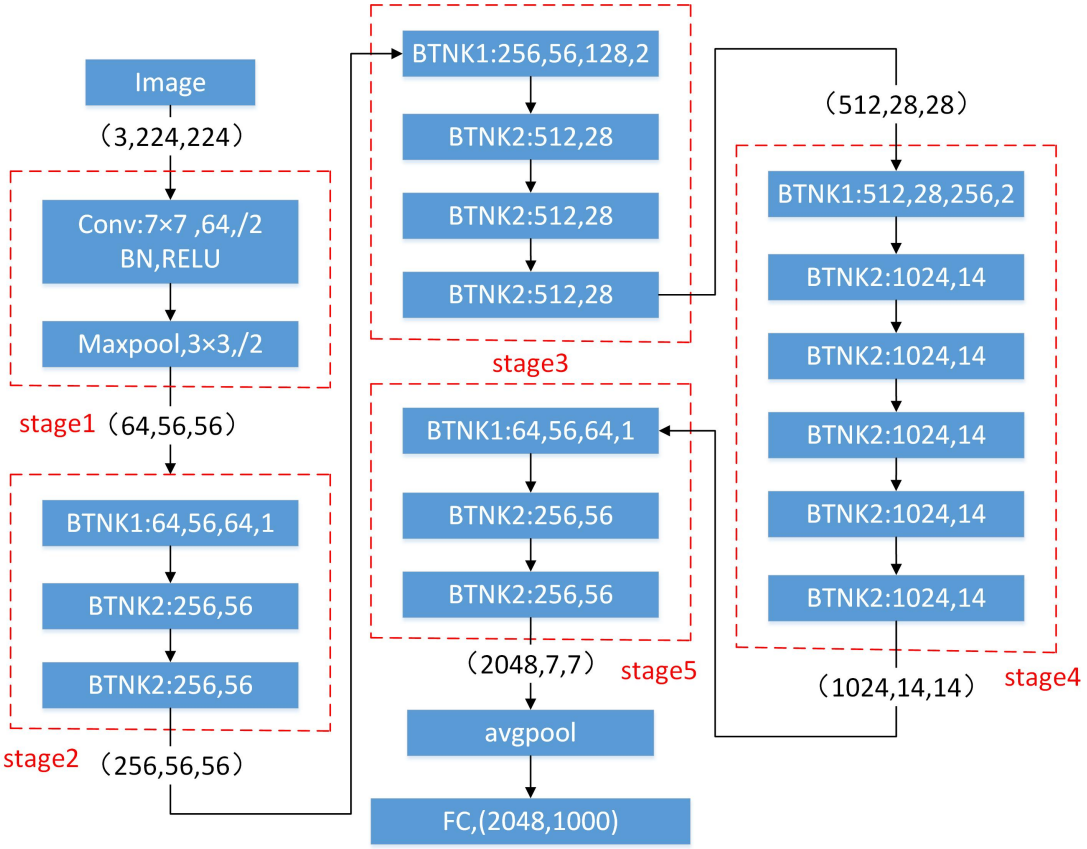


图 3-8 ResNet50 网络的整体框架

Bottleneck 分两种，一种输入和输出的通道数不同的 BTNK1（我们把 Bottleneck 用大写字母 BTNK 表示），另外一种输入和输出通道数相同的 BTNK2，如图 3-9 所示。BTNK1 的参数有 4 个，分别为 C，W，C1，S。BTNK2 的参数有 2 个，分别为 C，W。其中 C 和 C1 表示通道数，W 表示宽度和高度，S 表示步长。BTNK1 的右侧比 BTNK2 多了一个卷积层(设这个卷积操作函数为 $G(x)$)，我们设输入为 x，左边的 3 个卷积操作为 $F(x)$，输出为 $H(x)$,我们可以得到 BTNK1 的输出表达式为：

$$H(x)=F(x)+G(x) \tag{3-13}$$

BTNK2 的输出表达式为：

$$H(x)=F(x)+x \tag{3-14}$$

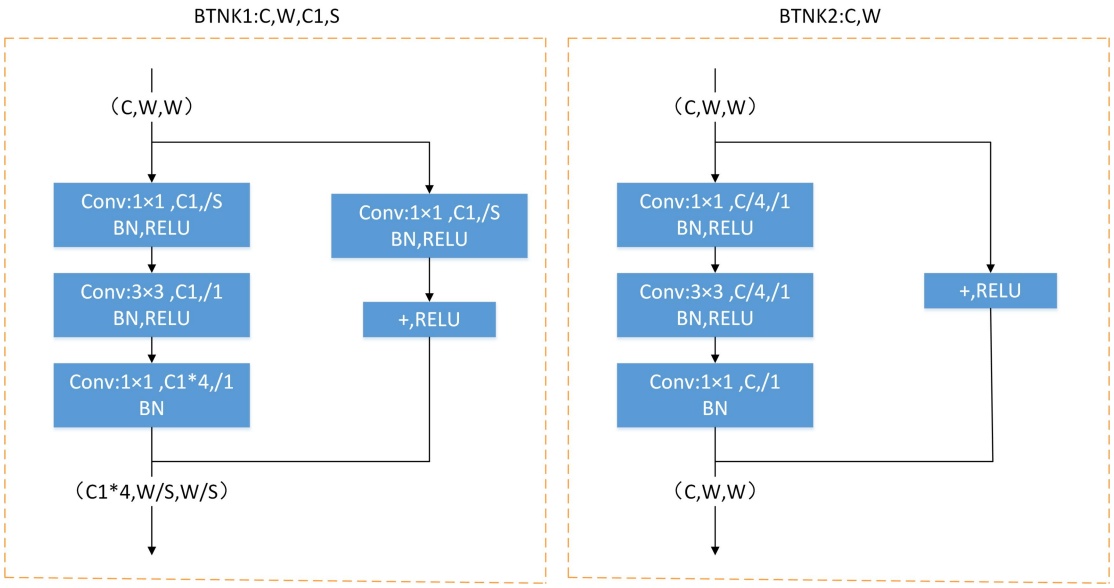


图 3-9 BTNK 的网络结构

注意力模块是一种即插即拔的网络模块，使用非常简单，一般加在卷积模块后面，针对 ResNet50 的特殊结构，为了使用 ResNet50 预训练的参数，我们不能将注意力模块加入到 block 里面去，因为这样会改变 ResNet50 的网络结构，从而导致不能使用预训练参数，因此我们选择在 ResNet50 网络第一阶段的第一个卷积层和第五阶段的最后一个卷积层的后面引入同一种注意力模块，分别为 SENet 模块，BAM 模块和 CBAM 模块三种注意力模块，添加后的 ResNet50 网络如图 3-10 所示。

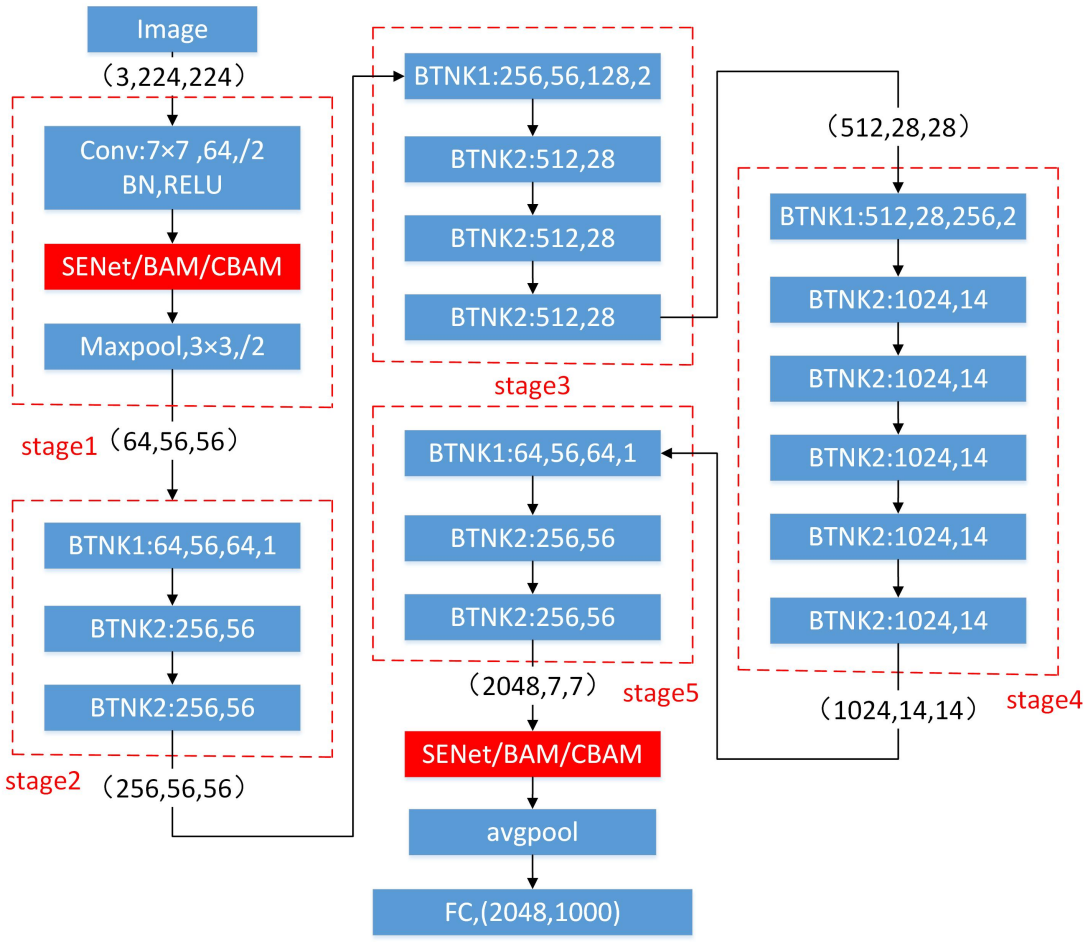


图 3-10　添加注意力模块 ResNet50 网络

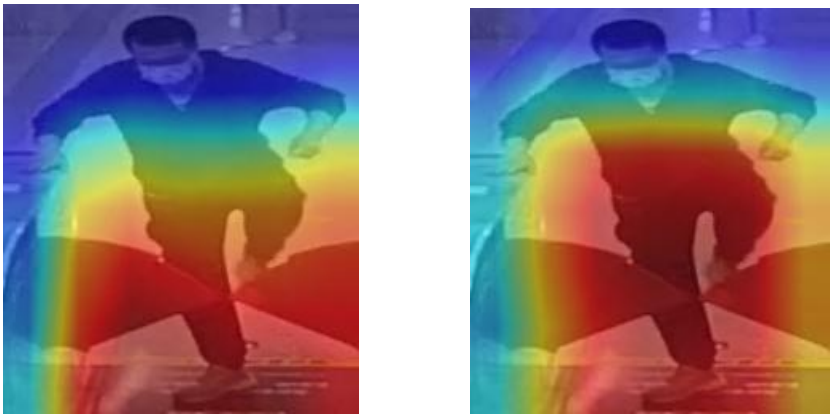
3.3.3　添加不同注意力模块实验结果对比分析

对加入注意力模块的 ResNet50 分别进行和前面训练方式一样进行训练，经过训练后的添加注意力模块的 ResNet50 网络对原图和骨架图的行为识别结果对比如表 3-5 所示。

表 3-5 ResNet50 添加不同注意力模块的预测结果对比

数据类型	模型	Precision(%)	Recall(%)	F1-Score(%)
Ori	ResNet50	82.843	81.759	82.297
Ori	ResNet50+SENet	83.943	82.536	83.234
Ori	ResNet50+BAM	83.010	82.331	82.669
Ori	ResNet50+CBAM	**84.395**	**82.945**	**83.663**
Skeleton	ResNet50	81.362	80.164	80.758
Skeleton	ResNet50+SENet	82.470	81.391	81.927
Skeleton	ResNet50+BAM	81.472	80.573	81.020
Skeleton	ResNet50+CBAM	**82.973**	**81.718**	**82.341**

由对比实验结果可知，针对 ResNet50 添加不同注意力模块后对非法过闸机行为识别数据集进行训练后的分类预测结果对比可以发现，注意力机制有助于提升网络的性能，添加 CBAM 模块的 ResNet50 网络在原始图像和骨架图像数据集上的分类效果最好，相对于没有加注意力模块的 ResNet50 网络在原图数据集上的精确率提升了 1.552%，召回率提升了 1.186%，F1-Score 提升了 1.366%。在骨架图像的数据集上添加 CBAM 模块的 ResNet50 网络相对于没有加注意力模块的 ResNet50 网络在精确率上提升了 1.611%，召回率提升了 1.554%，F1-Score 提升了 1.583%。SENet 注意力模块是只关注通道的注意力机制，而 CBAM 注意力模块不仅关注不同通道像素的重要性，还考虑到了同一通道不同位置像素的重要性，是一种结合通道和空间的注意力机制模块，相对于 SENet 注意力模块，CBAM 注意力模块在原图数据集上的精确率提升了 0.452%，召回率提升了 0.409%，F1-Score 提升了 0.429%。在骨架数据集上的精确率提升了 0.503%，召回率提升了 0.327%，F1-Score 提升了 0.414%。因此 CBAM 相比于只关注通道的 SENet 可以取得更好的效果。BAM 注意力模块虽然也结合了通道注意力和空间注意力，但是 BAM 模块是并联的结构，而 CBAM 模块是串联结构，相比于 BAM 结构更加简单，并且通道注意力是全局的，空间注意力是局部的，先关注全局再关注局部这种按顺序生成注意力图比并行生成注意力图更好。相对于 BAM 注意力模块，CBAM 注意力模块在原图数据集上的精确率提升了 1.385%，召回率提升了 0.614%，F1-Score 提升了 0.994%。在骨架数据集上的精确率提升了 1.501%，召回率提升了 1.145%，F1-Score 提升了 1.321%。因此 CBAM 相比于 BAM 可以取得更好的效果。由此可见，在 ResNet50 网络中添加注意力机制对非法过闸机行为识别分类效果有所提升，其中添加 CBAM 注意力模块对非法过闸机的识别效果最好。



(a) ResNet50　　(b) ResNet50+CBAM

图 3-11 热力图对比

为了更好地分析 CBAM 模块对模型性能的提升，本文使用热力图对比不加 CBAM 模块的 ResNet50 对非法过闸机行为的成像分布和在 ResNet50 中增加 CBAM 模块后对非法过闸机行为的成像分布。实验结果如图 3-11 所示。热力图是一种可以对卷积神经网络可视化的一种技术。它类似于红外成像图，温度越高成像越红，温度越低成像越蓝。同理，热力图通过权重的形式进行展现，卷积神经网络对图中的哪一部分越关注，成像越红。由图 3-11 可知，两个热力图都比较关注人体下半身的高抬腿动作，相对于没有添加 CBAM 模块的 ResNet50，添加 CBAM 模块的 ResNet50 更加关注人体的腿部动作，由此可见 CBAM 注意力模块对网络性能的提升是有效的。

3.5 本章小结

本章主要介绍了基于原图和骨架图的非法过闸机行为识别研究，首先使用不同的网络对非法过闸机行为识别数据集进行训练，测试不同网络对非法过闸机行为识别的分类效果对比，对比后发现 ResNet50 网络的分类效果更好，然后针对 ResNet50 模块引入注意力机制研究，探索引入 SENet，BAM，CBAM 三种不同注意力机制模块的 ResNet50 网络对非法过闸机行为识别数据集的分类效果。通过对比发现，针对非法过闸机行为识别数据集，添加 CBAM 模块的 ResNet50 网络分类效果最好。并使用热力图来证明 CBAM 模块对网络性能提升的有效性。

第 4 章 结合原图和骨架图的非法过闸机行为识别研究

4.1 研究动机

在上文我们已经做了基于原始图像和骨架图像的非法过闸机行为识别的研究，但是仅用一种原始图像数据或者骨架图像数据的信息做训练具有很大的局限性，因此本文打算对原始图像信息和骨架图像信息进行一个信息融合。基于信息融合的技术恰好可以弥补单一信息的缺点，可以很好地利用多种信息之间的互补信息来提高模型的表达能力，从而提升模型的识别能力。而信息融合的方式非常多，有基于像素级的前期融合方式，基于特征级的中期融合方式，基于决策级的后期融合方式，本文将基于非法过闸机原始图像信息和骨架图像信息进行信息融合的行为识别研究。

4.2 基于信息融合的非法过闸机行为识别

4.2.1 前期融合算法设计

前期融合是一种数据级融合方式，简单地说就是一种像素级别的融合。前期融合将多个独立的数据集融合成一个单一的特征向量，然后输入到机器学习分类器中。前期融合的优势在于流程比较简单，信息比较完整，但是前期融合的图像属于多通道叠加的图像，保留的都是数据的原始信息，而原始特征往往存在大量的冗余信息，对复杂数据而言，前期融合的意义并不大。我们的前期融合方式就是直接将非法过闸机行为识别数据集中的原始图片和对应的骨架图片上各点的像素值进行相加后取平均值操作，然后以这种方式获得像素级融合的数据作为训练样本，使用 ResNet50_CBAM 网络进行训练，同样对非法过闸机的数据和正常过闸机的数据按照 1:2 的比例进行训练，从而训练出一个基于前期融合的网络。然后使用原始图片和骨架图片做该网络的输入图片，进行前期融合网络的行为识别，最终获得分类识别结果。基于像素级的信息融合结构如图 4-1 所示。

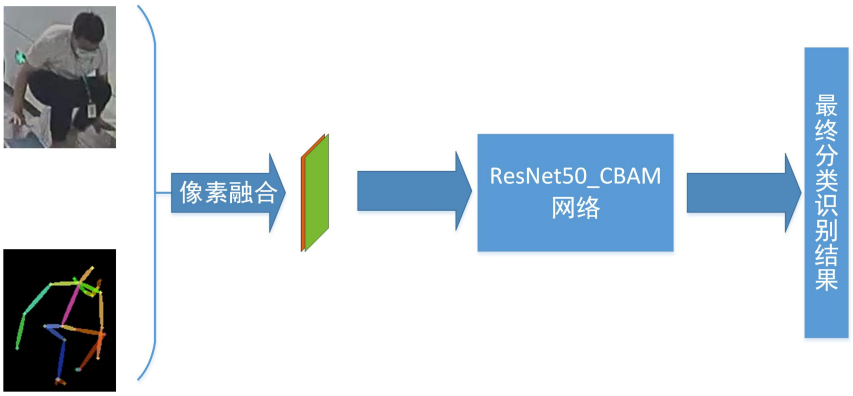


图 4-1 基于像素级的信息融合结构

4.2.2　中期融合算法设计

中期融合是一种基于特征级的融合方式，指将不同的数据通过一个深度神经网络先转化为高维特征向量，再于模型的中间层进行特征融合，最后进行目标的分类识别。以神经网络为例，中期融合首先利用神经网络将原始数据转化成高维特征表达，然后获取不同数据在高维空间上的共性。中期融合方法的一大优势是可以灵活的选择融合的位置。经过深度神经网络获得的特征具有维数多，尺度小的特点，容易在融合的时候出现特征分解匹配不对应，目标特征丢失等情况，因此，我们可以在中期融合时加入先验知识，从而保留目标特征。我们先使用 ResNet50_CBAM 对原始图像和骨架图像进行特征提取，然后将各网络最后输出的 2048 个特征向量进行拼接，从而实现特征级的融合。然后将该特征输入到 4096 层的全连接层进行下一步的处理，最终获得基于特征级融合的分类结果。基于特征级的信息融合结构如图 4-2 所示。

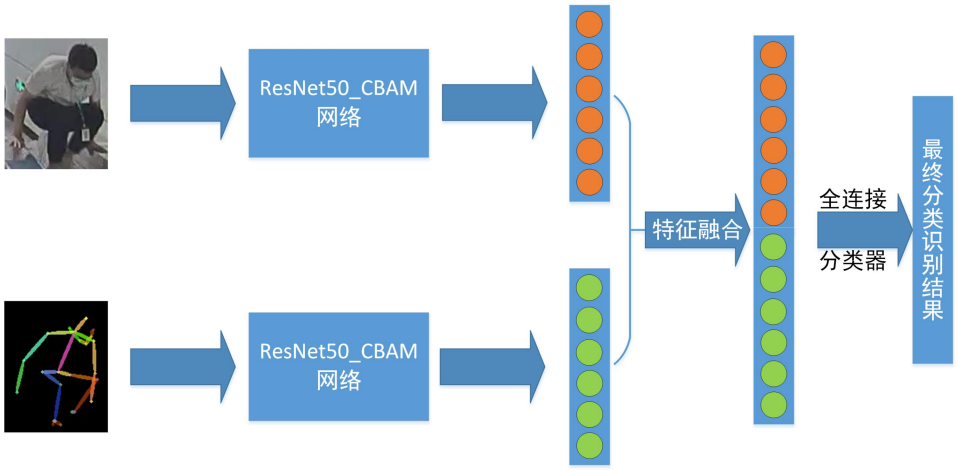


图 4-2　基于特征级的信息融合结构

4.2.3　后期融合算法设计

后期融合是一种基于决策级的融合方式，这种方式是将不同数据经过深度神经网络训练后，在分类器输出的分数上进行决策融合。后期融合则是将不同数据经过深度神经网络训练后，在分类器输出的分数上进行决策融合。后期融合的优势在于决策和数据的关联比较远，所以对数据源的依赖程度有所降低，分析能力得到了加强，并且融合模型的错误来自不同的分类器，而来自不同分类器的错误往往互不相关、互不影响，不会造成错误的进一步累加。但是因为每种网络分别训练一种数据，如果是多种数据源的情况下需要占用比较大的存储空间和计算空间。常见的后期融合方式包括最大值融合、平均值融合、贝叶斯规则融合以及集成学习等。其中集成学习作为后端融合方式的典型代表，被广泛应用于通信、计算机识别、语音识别等研究领域。同样我们对原始图像和对应骨架图像的非法过

闸机行为识别数据集分别用 ResNet50_CBAM 网络进行训练，在测试的时候让原始图像和骨架图像输入到对应的 ResNet50_CBAM 网络进行测试，我们可以获得两种数据的预测结果置信度，然后我们对这两种置信度进行决策融合算法，这里我对两种结果各占 0.5 的权重进行分配，也就是取两种置信度的平均值作为最终决策级的分类结果。最终置信度大于 0.5 则预测为非法过闸机行为。基于决策级的信息融合结构如图 4-3 所示。

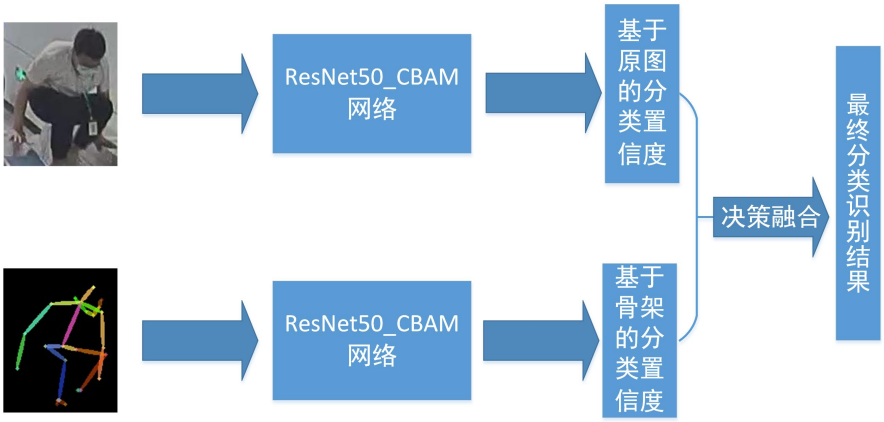


图 4-3 基于决策级的信息融合结构

4.2.4 不同信息融合方式实验结果对比

通过第三章对比实验分析可知，使用 ResNet50_CBAM 网络对非法过闸机行为识别的效果更好，所以本文使用 ResNet50_CBAM 作为信息融合网络的骨干网络，以原图和骨架图作为输入，将非法过闸机行为数据和正常过闸机数据集按照 1:2 的比例进行训练，训练周期为 50 个 epoch，训练的 batch size 为 128，学习率设为 0.0008，上一次梯度更新的权重 momentum 设为 0.9，权重损失 weight decay 设为 1e-4。分别使用上文所说的前期融合，中期融合，后期融合三种融合方式在非法过闸机行为识别数据集上按照非法行为与正常行为进行 1:2 的比例训练，在测试集上的预测效果如表 4-1 所示。

表 4-1 不同信息融合方式预测结果对比

模型	Precision(%)	Recall(%)	F1-Score(%)	单帧图片平均处理时间（ms/帧）	Model Size
ResNet50+CBAM	84.395	82.945	83.663	**59**	**215M**
EarlyMergeModel	84.542	82.986	83.756	63	220M
MiddleMergeModel	**87.806**	**85.112**	**86.438**	76	438M
LastMergeMode	85.774	84.090	84.924	71	430M

如表 4-1 所示，通过基于原始图像的最好结果与不同信息融合方式的对比实验结果分析可知，信息融合网络比基于原始图像分类有更好的分类效果，但是基于原始图片的处理速度更快，模型更轻。但是相对来讲在损失少许时间的情况下分类效果的提升更加重要。在信息融合实验对比中，使用中期的特征级融合方式对非法过闸机行为识别的效果最好，而前期的像素级和后期的决策级融合方式对非法过闸机行为识别的效果相对于原始图像分类效果也有所提升，我们设计的前期融合的图像属于像素值上的加权平均图像，保留的都是数据的原始信息，而原始特征往往存在大量的冗余信息，所以前期融合的效果提升还不够明显。而后期决策级融合方式属于对分类结果上的进一步优化，实际融合的信息并不多，并没有充分利用两种信息间的共性信息，因此对非法过闸机行为识别的效果提升也不够好。而中期的特征级融合可以很好得利用两种信息间的共性知识进行关键特征的提取，进行更好的融合，因此中期特征融合的方式效果更好。

4.3 基于 OSFANet 网络的行为识别

4.3.1 OSFANet 网络框架设计

由上文中可知前期融合，中期融合，和晚期融合三种信息融合方式对非法过闸机行为识别的分类效果均有所提升，其中基于特征级的融合方式最好。基于前面信息融合的初步实验，本文提出了一种添加注意力机制的信息融合网络（Ori and Skeleton Fusion With Attention Net, OSFANet），该网络的框架如图 4-4 所示。该框架由三个分支组合而成：原始图像分支，骨架图像分支以及结合原图和骨架图信息的超分支。原始图像分支主要是用来提取原始图像中的特征，骨架图像分支主要是对原图中对应的骨架信息进行特征提取，这两个又可以称为主分支，而超分支主要是对主分支中 ResNet 网络的第二阶段和第三阶段的两个主分支所提取到的特征进行整理，结合与完善，并输入到超分支中的 OSFA 模块进行处理。从图 4-4 中可知，本文设计的 OSFA 网络之所以选择对主干网络的第二阶段之后的特征进行提取并融合的主要原因是因为第一阶段的特征处于图像预处理阶段，这个阶段的特征比较浅并且多，在经过第二阶段之后的特征就属于深层次的关键特征，在这个时候进行融合和在每个阶段都融合相比减少了计算量，并且融合的特征质量更高，减少了模型过拟合的风险。在两个主分支和一个超分支的共同作用下，我们得到了三个预测结果的损失函数以及预测结果置信度，我们的模型总损失函数 *Loss* 被定义为：

$$Loss = \alpha_1 * loss_O + \alpha_2 * loss_S + \alpha_3 * loss_H \tag{4-1}$$

其中 $Loss_O$ 和 $Loss_S$ 分别表示原图和骨架图对应的主分支损失，$Loss_H$ 表示融合特征的超分支损失，其中 α_1 α_2 α_3 表示各分支损失函数对应的权重比，α_1 设为

0.35，α_2 设为 0.25，α_3 设为 0.4。

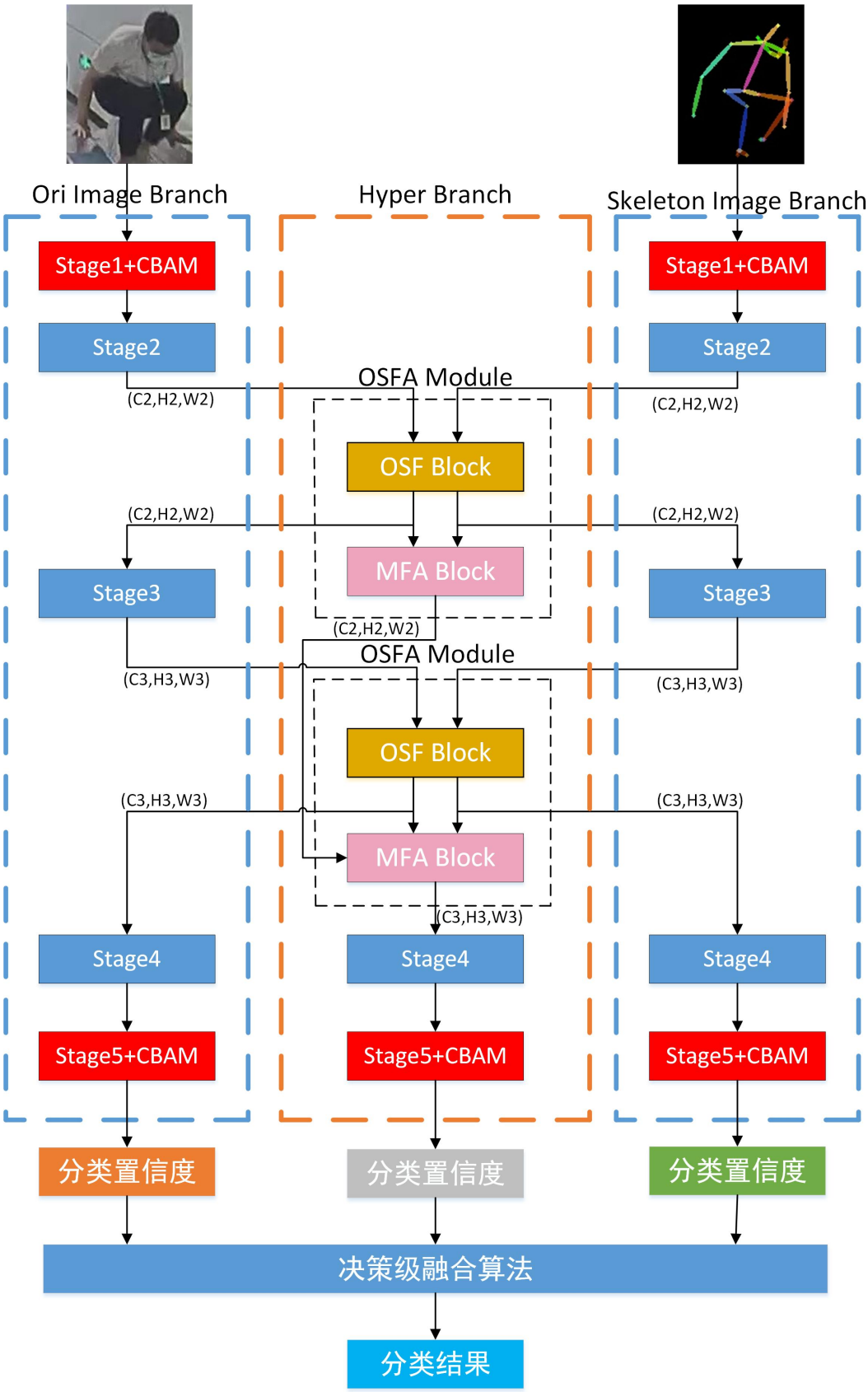


图 4-4 OSFANet 网络框架

针对三个分支输出的预测结果置信度，我们设计的决策级融合算法如下式4-2 所示，我们设原图和骨架图的主分支输出的置信度为F_O和F_S，超分支的置信度为F_H，经过决策级融合的置信度为F，最终我们以决策融合之后输出的置信度F作为判断是否存在非法过闸机行为的标准，如果F大于 0.5，则预测为非法过闸机行为。其中α_1 α_2 α_3表示各置信度对应的权重比，α_1设为 0.35，α_2设为 0.25，α_3设为 0.4。

$$F = \alpha_1 * F_O + \alpha_2 * F_S + \alpha_3 * F_H \tag{4-2}$$

4.3.2 OSFANet 主干网络设计

由上文已经知道 ResNet50 网络对非法过闸机行为识别数据集的分类效果最好，并且在以融合CBAM 注意力模块后的 ResNet50 对非法过闸机数据集的分类效果有所提升，因此本文选取融合 CBAM 注意力模块的 ResNet50 网络作为原图和骨架图输入的主干网络。

由上文介绍已知 ResNet50 网络每个阶段的输出信息作为下一个阶段的输入信息，而本文中的 OSFANet 需要融合两个主干网络的第二阶段和第三阶段的特征，因此我们需要把原 ResNet 网络中的第二阶段到第三阶段和第三阶段到第四阶段的连接断开，将原图和骨架图经 ResNet 网络第二阶段的输出特征（设为C2×H2×W2）一起输入到超分支中的 OSFA 模块中，OSFA 模块可以对不同信息的特征进行融合学习并找出信息间的相关性，经过 OSFA 模块指导优化后的信息作为 ResNet 网络第三阶段的输入（C2×H2×W2），经过第三阶段输出后（设为C3×H3×W3）的特征再次输入到 OSFA 模块进行融合学习，经 OSFA 模块指导优化后再次输入（C3×H3×W3）到 ResNet 网络的第四阶段，最终使得两个主干网络的输入信息不再相互独立，而是一个可以通过超分支进行互相学习的网络。最终提升了主干网络的特征表达能力。

超分支作为连接两种信息的桥梁放在原图和骨架图输入的两个主分支之间，由两个 OSFA 模块组成，其中第一个 OSFA 模块有两个输入，接收两个主分支第二阶段的输出信息。第二个 OSFA 模块有三个输入，接收第三阶段的输出信息以及第一个 OSFA 模块的输出信息。输入到 OSFA 模块的信息经过互相学习指导后输出到原主分支的下一个阶段，经过 OSFA 模块融合的信息输出到下一个 OSFA 模块，最终在第二个 OSFA 模块输出经过特征融合后的信息。该信息是两种数据在高维特征空间的一个共性表达，从而使输入数据的信息范围更宽，提高了网络的性能。

4.3.3 超分支 OSFA 模块设计

原始图像和骨架图像可以从不同方层面来表现非法过闸机行为，并且这两种

信息之间存在着共性和互补的信息。而目前基于前期融合，中期融合和晚期融合的信息融合技术主要是专注于融合之后的信息，却没有对互补的信息进行关注。本文设计的 OSFA 模块由原图和骨架图融合模块 OSF（Ori And Skeleton Fusion）和多尺度注意力融合模块 MFA（Mutiscale Fusion With Attention）共同实现。该模块不仅能融合主分支中第二阶段和第三阶段的输出信息，还能够自适应学习两种数据之间的互补信息。以此来优化和完善原始图像和骨架图像中特征的提取。

4.3.3.1 OSF 模块

OSF 模块的主要作用是对原图和骨架图这两种数据的信息进行提取并学习两种数据的互补信息，最后将学习到的互补信息输入到主分支中的下一个阶段作为一种监督信息。从而达到优化主分支的特征提取能力。如图 4-5 所示。

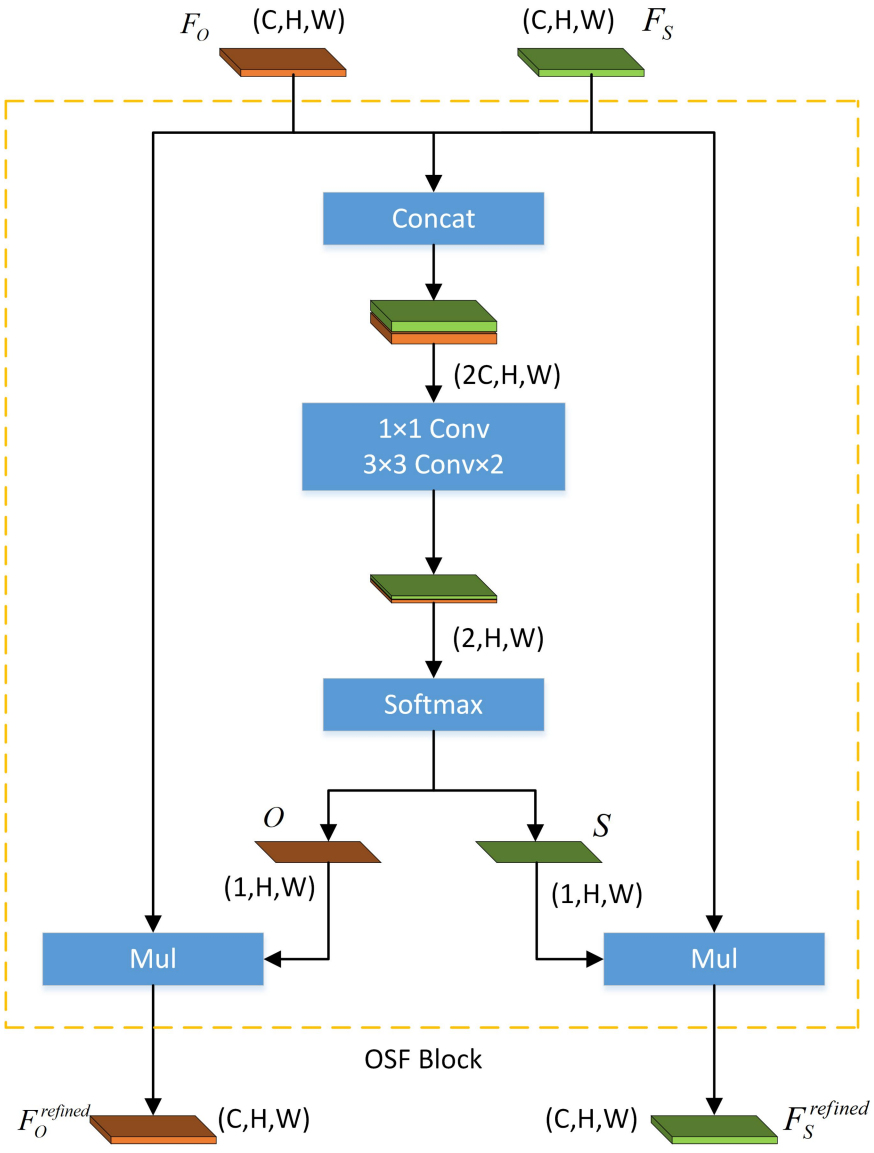


图 4-5　原图和骨架图融合模块

我们首先将该网络中两个主分支输出的原图特征 F_O 和骨架特征 F_S 进行拼接，假设输入的特征图大小为 C×H×W（其中 C 表示通道数，H 表示高度，W 表示宽度），我们使用 1×1 的卷积对输入的特征进行信息交互整合，同时将通道数减少为 C/2，以此来实现信息的压缩，减少卷积运算。然后再输入到两个 3×3 的卷积层中，第一个卷积将通道数减少为 C/4,第二个卷积将通道数变成 1。为了提高超分支的非线性表达能力，在两个卷积层后面均添加了 ReLU 激活函数。在以上三个卷积的作用下，主分支的特征图的通道数被压缩为 1 层，视作原主分支的注意力图谱。然后将两个注意力图谱在通道维度上进行串联，形成两通道特征图，再使用 softmax 函数进行两通道特征图的交叉注意力整合，如下式：

$$O_i = \frac{e^{O_i}}{e^{O_i} + e^{S_i}}, i = [1,2,3,\cdots, H \times W] \tag{4-3}$$

$$S_i = \frac{e^{S}}{e^{O_i} + e^{S_i}}, i = [1,2,3,\cdots, H \times W] \tag{4-4}$$

其中，O_i，S_i 代表原图和骨架图的特征图对应位置 i 处的注意力值，经过 softmax 输出的注意力图谱 O 和 S 不仅包含着原始信息，还包含着另一个主分支的信息，并且可以自适应调整两种信息的重要程度。通过这种方式输出的 O 和 S 可以作为原主分支信息的注意力权重来优化原主分支的输入特征，如下式：

$$F_O^{refined} = O \bullet F_O \tag{4-5}$$

$$F_S^{refined} = S \bullet F_S \tag{4-6}$$

其中，$\bullet$ 表示点乘，也就是特征向量中的对应元素相乘。$F_O^{refined}$ 和 $F_S^{refined}$ 表示经过 OSF 模块互补学习后的优化的主分支特征。综上所述，OSF 模块可以看作是一种空间注意力模块，它可以在不改变输入特征的像素关系和空间的情况下，对主分支的输入特征进行信息交互和优化，从而提高主分支的特征表达能力。

4.3.3.2 MFA 模块

MFA 模块不同于 OSF 模块，OSF 模块主要作用是对主分支的输入特征进行调整和优化，而 MFA 模块主要是对 OSF 模块输出的两个主分支特征以及前一个 OSFA 模块输出的融合特征进行多尺度融合，以此来获取不同特征中的交叉信息，从而实现融合特征的提取。如图 4-6 所示，MFA 模块的输入由经过 OSF 模块输出的两个主分支特征 $F_O^{refined}$ 和 $F_S^{refined}$ 和一个经过前一个 OSFA 模块输出的融合特征 F_{fusion} 三个部分组成，$F_O^{refined}$ 和 $F_S^{refined}$ 的大小和经过主分支残差块输出的 F_O 和 F_S 相同，我们设经过 OSF 模块输出的两个主分支的输入大小为 C×H×W，F_{fusion} 来自上一个 OSFA 模块的输出，并没有经过主分支的残差块，所以大小为 C/2×2H×2W，首先对两个主分支特征 $F_O^{refined}$ 和 $F_S^{refined}$ 进行矩阵上的加法运算，然

后让经过前一个 OSFA 模块输出的融合特征 F_{fusion} 输入残差块进行进一步的特征提取（如果是第一个 OBAF 模块，则没有这个步骤，而是将两个主分支特征 $F_O^{refined}$ 和 $F_S^{refined}$ 进行矩阵上的加法运算后直接接入 CBAM 注意力模块，最后输出为 F_{fusion}），这个残差块和 F_O 和 F_S 经过的残差块相同，从而使 F_{fusion} 经过残差块后输出的空间结构和主分支特征 $F_O^{refined}$ 和 $F_S^{refined}$ 保持一致。因为两个主分支特征加法运算后依旧是 C×H×W，所以它们进行拼接后的特征图大小为 2C×H×W，然后使用 3×3 的卷积层对拼接后的特征进行提取，并将其通道数从 2C 变成 C。为了使网络更加专注于感兴趣区域，再输入到 CBAM 注意力模块进行特征提取，最后获得多尺度融合的特征 F_{fusion}。

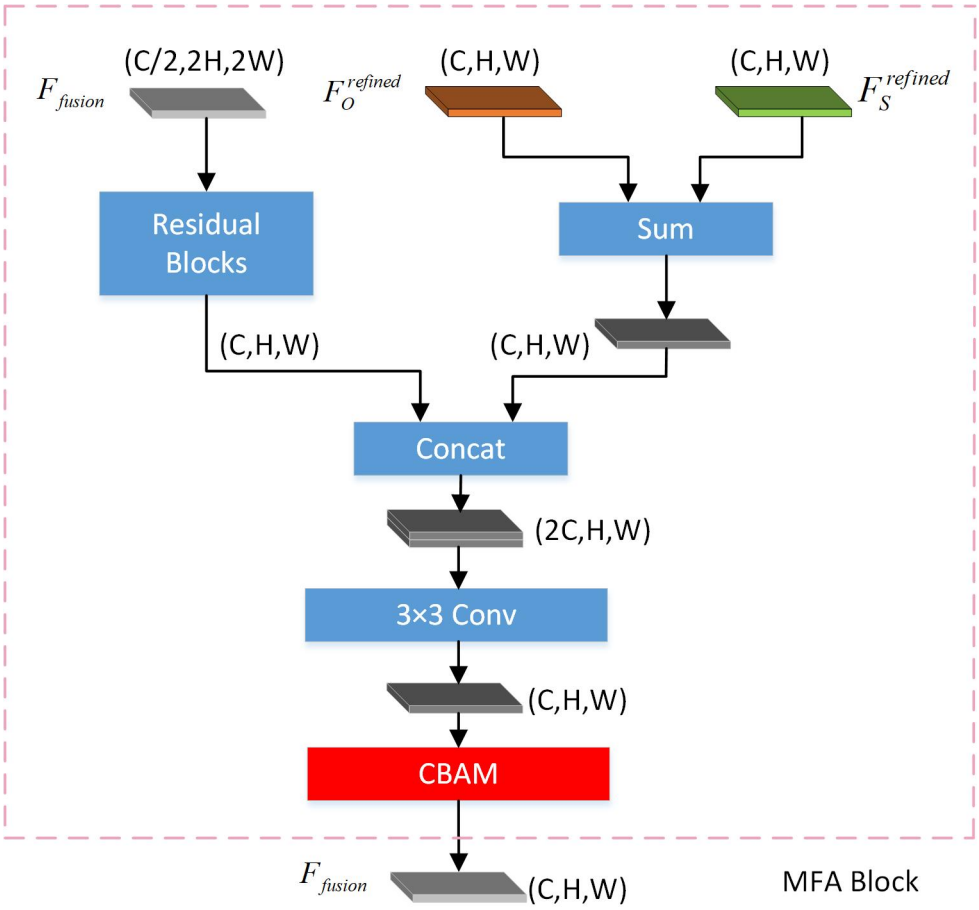


图 4-6　多尺度融合模块

4.3.4 实验结果分析

我们的训练方法和上文保持一致，将非法过闸机行为数据和正常过闸机数据集按照 1：2 的比例进行训练，训练周期为 50 个 epoch，训练的 batch size 为 128，学习率设为 0.0008，上一次梯度更新的权重 momentum 设为 0.9，权重损失 weight decay 设为 1e-4。OSFANet 网络模型训练好之后，使用测试集数据进行测试，在 2445 张非法过闸机行为的图片中，识别为非法过闸机行为的图片有 2366 张，79

张误检为正常过闸机行为。在 5839 张正常过闸机行为图片中，5517 张图片识别为正常过闸机行为，322 张图片识别成非法过闸机行为。混淆矩阵如表 4-2 所示。

表 4-2 混淆矩阵分类结果

预测结果\实际结果	非法过闸机	正常过闸机
非法过闸机	2140	233
正常过闸机	305	5606

通过表我们可以计算出 OSFANet 的精确率为 90.181%，召回率为 87.526%。F1-Score 为 88.834%。为了验证 OSFANet 模块的有效性，和上文中的信息融合网络进行实验对比，预测结果对比如下表 4-3 所示。

表 4-3 不同信息融合网络预测结果对比

模型	Precision(%)	Recall(%)	F1-Score(%)
ResNet50+CBAM	84.395	82.945	83.663
EarlyMergeModel	84.542	82.986	83.756
MiddleMergeModel	87.806	85.112	86.438
LastMergeMode	85.774	84.090	84.924
OSFANet	**90.181**	**87.526**	**88.834**

由表可知，本文提出的 OSFANet 网络在非法过闸机行为识别测试集上的分类效果最好，比中期融合网络的精确率提升 2.375%，召回率上提升 2.414%，F1-Score 提升 2.396%。相对于比 ResNet50_CBAM 网络的精确率提升 5.786%，召回率上提升 4.581%，F1-Score 提升 5.171%。由此可见，针对非法过闸机行为识别数据集，本文提出的 OSFANet 网络更好，因此，我们选择本文提出的 OSFANet 网络作为非法过闸机行为识别的模型。

4.3.5 消融实验分析

OSFANet 网络的独特之处在于网络中超分支里面的特征融合模块 OSFA 和最后三个分支输出时使用的决策融合模块。这个 OSFA 模块主要由两部分组成：OSF 模块和 MFA 模块。在这两个模块的共同作用下，OSFA 模块不仅可以挖掘两个主分支之间的互补信息并对主分支特征进行优化调整，还可以对两个特征进行融合。而对三分支输出采用决策融合方式使预测结果更加精细化。为了验证 OSF 模块，MFA 模块和三分支输出采用决策融模块的有效性，我们进行了消融实验，分别去除其中一个模块来比较消融后网络的性能。

为了验证 OSF 模块的有效性，本文去除了 OSFA 模块中的 OSF 模块，对主分支中 ResNet 网络的各个阶段保持原始 ResNet 网络的连接方式，并且将两个主

分支中第二阶段和第三阶段的输出同时引入 OSFA 模块中的 MFA 模块，也就是直接将原始图片和骨架图片的特征直接作为 MFA 模块的输入，其余地方保持不变进行消融实验。

为了验证 MFA 模块的有效性，本文去除了 OSFA 模块中的 MFA 模块，保持 OSF 模块的正常使用，相当于少了第一个 OSFA 模块的融合特征与第二个 OSFA 模块的信息进行多尺度融合的过程，取而代之的是直接将主分支第三阶段的输出特征在 OSF 模块优化后进行加权平均操作后直接输入到超分支的第四阶段，从而实现消融实验。

为了验证决策融合模块的有效性，本文去除了 OSFANet 的决策融合模块，取而代之的是直接使用超分支的预测结果作为 OSFANet 的输出结果，从而实现消融实验。

在进行三种方式的消融实验后，我们的实验结果如表 4-4 所示。

表 4-4 不同消融方式实验结果对比

模型	Precision(%)	Recall(%)	F1-Score(%)
Without OSF	88.077	85.808	86.928
Without MFA	88.692	85.971	87.310
Without Decision	89.487	86.339	87.885
OSFANet	**90.181**	**87.526**	**88.834**

从表中可知，相比于 OSFANet 网络，在去除 OSF 模块后，非法过闸机行为识别的精确率下降了 2.104%，召回率下降了 1.718%，F1-Score 下降了 1.906%。因为在去除了 OSFA 模块中的 OSF 模块后，对主分支中 ResNet 网络的各个阶段保持原始 ResNet 网络的连接方式，并且将两个主分支中第二阶段和第三阶段的输出同时引入 OSFA 模块中的 MFA 模块，也就是直接将原始图片和骨架图片的特征直接作为 MFA 模块的输入，这样就使得两种信息不能进行优化调整从而获得互补信息。由此可见，OSF 模块使两种互补信息进行优化调整是有效的。在去除 MFA 模块后，精确率下降了 1.489%，召回率下降了 1.555%，F1-Score 下降了 1.524%。因为只保持 OSF 模块的正常使用，相当于少了第一个 OSFA 模块的融合特征与第二个 OSFA 模块的信息进行多尺度融合的过程，取而代之的是直接将主分支第三阶段的输出特征在 OSF 模块优化后进行加权平均操作后直接输入到超分支的第四阶段，这样就导致主分支信息缺少多尺度融合的过程。由此可见，MFA 模块对主分支信息进行多尺度融合是有效的。在去除决策融合模块后，精确率下降了 0.694%，召回率下降了 1.187%，F1-Score 下降了 0.949%。因为直接使用超分支的预测结果作为 OSFANet 的输出结果，缺少两个主分支的预测信息

进行加权的过程。由此可见，决策融合模块可以让模型的预测结果更加精细，从而提升分类效果是有效的。

4.4 本章小结

本文主要是研究结合原图和骨架图对非法过闸机行为识别的分类效果。先设计了基于像素级的前期融合方式，基于特征级的中期融合方式和基于决策级的后期融合方式进行不同时期融对原始图像和骨架图像进行融合的实验，经过实验对比发现前期融合和后期融合对非法过闸机行为识别的分类效果提升不够明显，基于特征级的中期融合方式的分类效果优于前期融合和后期融合。最后本文设计一种添加注意力机制的 OSFANet 网络，相较于简单的中期融合和基于原图的 ResNet50_CBAM 网络有更好的分类效果，并做了消融实验验证 OSFANet 的有效性。因此，选择本文提出的 OSFANet 网络作为非法过闸机行为识别的模型。

第 5 章 基于视频行为事件的非法过闸机行为检测系统设计

5.1 研究动机

随着以卷积神经网络为代表的深度学习的快速发展。基于单张图片的方法进行行人行为分析已广泛应用于现实生活中，给人们的生产生活带来了极大的便捷。其中，目标检测算法已经广泛用于生活场景图像检测与识别；骨架检测技术也同样开始广泛应用于人的行为和姿态的分析识别。但是，当前的目标检测和骨架检测技术大都是基于单帧图片的，正如上章的非法过闸机行为识别算法，也是基于单帧图片的，无法从视频中的一系列连续动作判断人员的行为，直接应用于非法过闸机的行为检测是非常武断的，非常容易出现误报的情况。因此我们需要结合时间维度的三维信息来综合评定过闸机的人员是否存在非法过闸机行为。

5.2 基于视频的非法过闸机行为检测算法的设计

本文设计的非法过闸机行为检测算法从技术的角度来讲可以分为四个部分：视频预处理，行人检测，行为识别和行为分析。视频预处理是对视频的一个解析过程，是基于单帧图片算法实现的基本条件。行人检测技术是研究非法过闸机行为检测的第一步，只有先检测到人的存在才能开始对进行下一步的行为识别和行为分析，行为识别算法是本研究的核心算法，该算法是非法过闸机行为检测技术能否实现的关键技术，行为分析是对多帧结果的一个总体判断，使用相应的跟踪算法来实现对每个过闸机人员的行为判断。该算法的整体框图如图 5-1 所示。

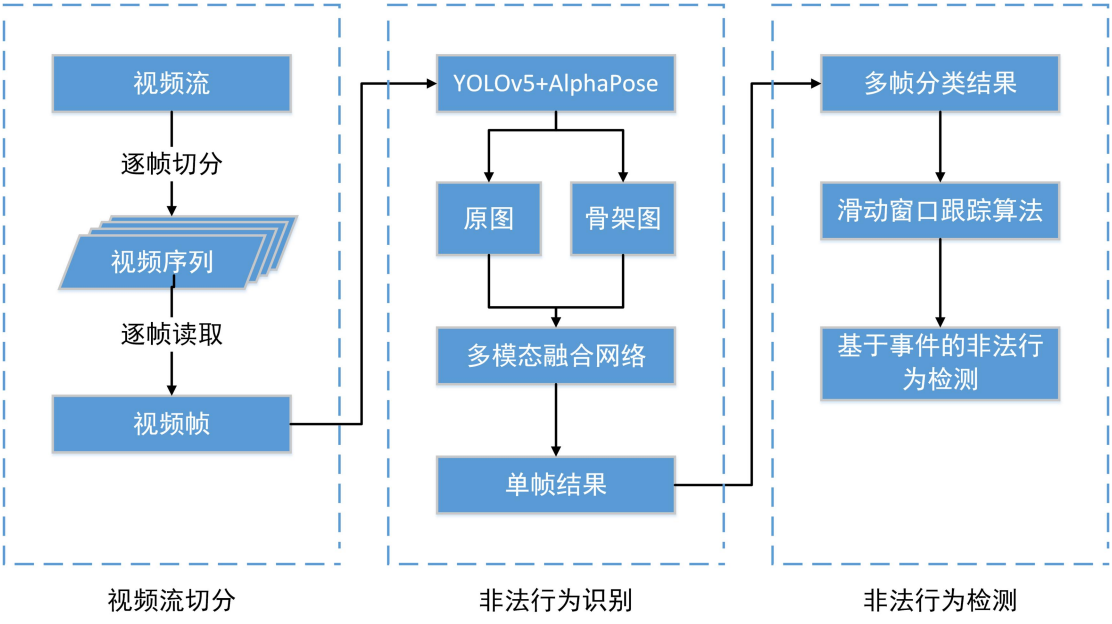


图 5-1 非法过闸机行为检测算法整体框图

由整体框图可以知道非法过闸机行为检测算法可以分为三个步骤：

步骤一：针对地铁闸机口的监控视频流，首先对将视频流切分为单帧的图像序列，依据输入的图像信息获取图像中闸机口的位置坐标，并划分一个包围闸机口的四边形区域作为待检测的区域。

步骤二：通过 YOLOv5 目标检测算法和 AlphaPose 骨架提取算法，获取待检测区域的行人信息和骨架信息，并构行人信息。再基于行人信息，构建基于 RGB 图像块和骨架结构图的信息识别网络，对生成行人信息进行分类识别，并给出相应的识别结果。

步骤三：基于单帧的行人信息和识别结果信息，使用滑动窗口对视频流中的的人员进行跟踪和行为分析，在滑动窗口内如果出现多次非法过闸机行为则对该人员报警，使用红色框进行框出并以图片的方式保存在报警文件夹。

针对非法过闸机行为检测算法的设计，本文在下面进行更加详细的介绍。

5.2.1 基于区域入侵的行人检测

5.2.1.1 闸机口监控视频要求

针对非法过闸机行为检测算法的设计，对地铁闸机口的监控视频有以下要求：

（1）视频流的分辨率为 1920×1080，视频格式为 mp4，视频的长度不得小于 5s；

（2）需要事先标记出视频流中地铁闸机口的位置；

（3）以分辨率为 1920×1080 的视频流为例，待检测人员的高不得小于 190 个像素，宽不得小于 56 个像素

（4）摄像头安装高度：摄像头需要能从上往下拍摄到所有完整的闸机区域；摄像头安装角度：正对闸机区域，左右角度偏移范围在 15 度以内，俯角控制在 30 度到 45 度之间；摄像头安装位置：摄像头安装位置与闸机靠近摄像头的一侧的距离应在 2 米至 5 米之间；

（5）其他要求：地铁进口处，摄像头需在闸机内部；地铁出口处，摄像头需在闸机外部，示意图如下图 5-2 所示。

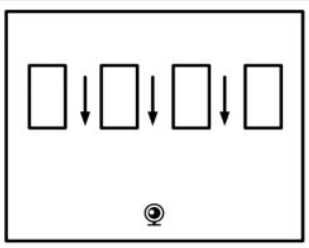


图 5-2 摄像头位置示意图

针对非法过闸机行为检测算法的设计，我们对地铁闸机口的监控视频的序列

采集步骤如下：

步骤一：对于输入的地铁闸机监控视频流，对视频流进行逐帧截取以获取相应的图像序列；

步骤二：考虑到行为发生的时间以及算力限制，将步骤一中的图像序列进行抽样，具体表现为连 5 帧选取一张，作为该 5 张图像的代表帧，并据此构建降采样后的图像序列用于后期识别；

步骤三：对于步骤一中获取的降采样图像序列，依据输入的地铁闸机口位置坐标，从图像中划分相应的待检测感兴趣区域。

5.2.1.2 区域入侵检测算法设计

在制作数据集那章已经讲到，由目标检测的研究现状可知 YOLOv5 算法在实时性和准确性上都有很大的优势，本研究数据集的制作以及非法过闸机行为检测算法中的行人检测算法都是 YOLOv5 目标检测算法，而本研究只对视频图片中划分的待检测区域四边形中的人员进行行人检测。为了实现待检测区域内的行人检测，我们需要使用相应的入侵检测[87]算法对行人进行是否在区域内的判断，目前常用的入侵判断方法主要有越界判断和和区域入侵检测这两种。

越界判断顾名思义就是人员在越过警戒线时进行判断，如果有行人进入待检测区域，则该人员的运动轨迹所形成的线段和我们所划分的四边形区域边界线段必定有交点，反正如果没有入侵则不会出现交点。如图 5-3 所示为越界判断图。

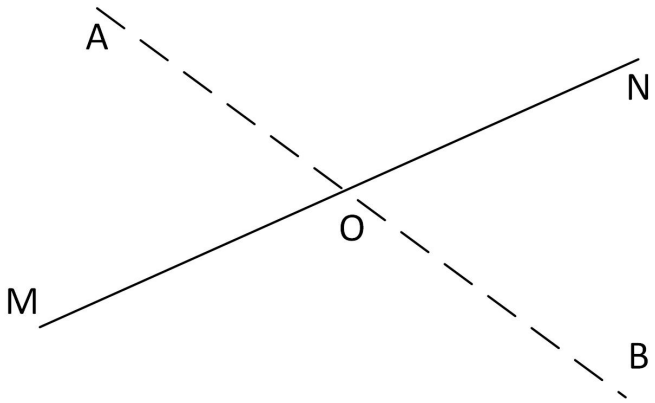


图 5-3 越界判断图

由图可知，实线 MN 表示待检测区域的四边形边界之一，虚线 AB 表示行人从四边形边界 MN 的外侧走向四边形边界 MN 内测的路径，当行人穿越边界 MN 时，虚线 AB 和实线 MN 一定会存在交点 O，如果行人没有进行越界行为，则行人的轨迹 AB 会在实线 MN 的同一侧，这种情况是没有交点的。我们设 M 点坐标为(x_M, y_M)，N 点坐标为(x_N, y_N)，A 点坐标为(x_A, y_A)，B 点坐标为(x_B, y_B)，设两条线交点 O 为(x, y)，我们可以把人员是否出现越界问题转化为二维平面中

的两条线段是否存在交点问题，则实线与虚线的交点可由式(5-1)表示：

$$\begin{aligned}\frac{y-y_M}{y_N-y_M}&=\frac{x-x_M}{x_N-x_M}\\ \frac{y-y_A}{y_B-y_A}&=\frac{x-x_A}{x_B-x_A}\end{aligned}\tag{5-1}$$

判断二维平面的两条线是否存在交点常用的方法有两种。第一种方法是先计算两条线段所对应的直线的交点，再判断该点是否再两条线段上，如果判断该交点在两条线段中任意一条线段上，那说明这两条线段之间存在交点，出现了越界行为，如果该交点没有在任意一条线段上，则说明这两条线段没有交点，也代表没有出现越界行为。第二种方法是先求出其中一条线段的两个端点，再判断这两个端点是否在另一条线段的同一侧，如果在同一侧则说明两条线段不存在交点，即没有发生越界行为，反之出现越界行为。第一种方法的原理比较简单易懂，求出两条直线的相交点，只要两个公式就可以得到结果。第二种方法相对复杂，需要以两条线段为对角线，根据对角线画出对应的矩形，根据矩形的重合情况来分多种情形进行讨论，计算量较高并且耗时，从而影响性能，本文使用第一种方法判断两线段是否相交。

区域入侵检测就是划定一片区域，判断目标的运动轨迹是否经过区域内。如果经过区域内则发生区域入侵行为，如果没经过则判断没有发生区域入侵行为。区域入侵检测示意图如图 5-4 所示。四边形 ABCD 为划分的待检测区域，虚线 P 为行人的运动轨迹，我们只需要判断点 P 的位置是否在四边形 ABCD 内部就可以判断是否出现了区域入侵。如果点 P 在四边形内部，则说明存在区域入侵，反之则不存在区域入侵行为。

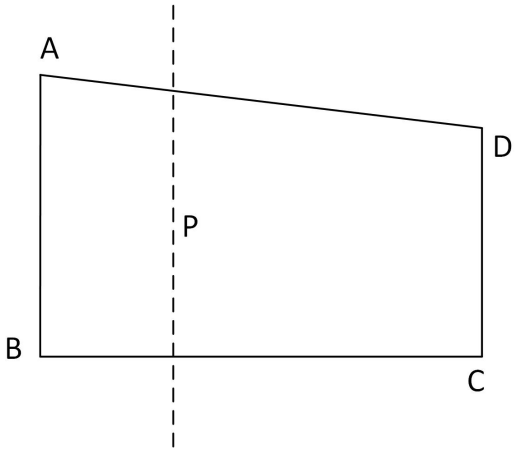


图 5-4 区域入侵检测示意图

由上图可知，我们只需要计算行人 P 是否在区域内就可以判断是否发生区域入侵行为。目前计算点是否在多边形内较多的方法有三种，第一种是射线法，顾

名思义就是以该点为端点，朝任意方向（例如朝 y 轴方向）作一条射线，如果越过多边形的边数是偶数，则说明该点在多边形之外，如果越过的边数是奇数，则证明该点在多边形之内。换言之就是求该射线与多边形的交点个数。第二种方法是称为弧长法，以该点为圆心作单位圆，然后计算多边形各边在该圆上的径向投影，然后计算所有边的径向投影代数和。如果和为2π，则说明点在多边形内，如果和为π，则说明点在多边形边上，如果和为 0，则说明点在多边形外。弧长法可以适用于多种情况而且准确，但是计算很耗时。第三种方法是使用图像处理工具生成一张多边形区域内全黑外部全白的图片，然后判断该点的像素值，若像素值为 255，则可以判断该点在多边形之内，如果像素值未 0，则该点在多边形之外。综上所述，第一种方法思路简单计算方便，第二种计算过于复杂，第三种不适合非法过闸机行为检测的后续处理，因此本文选用第一种方法来实现区域入侵检测。

针对进入待检测区域的行人，我们通过 YOLOv5 获得该人员的行人框位置坐标，判断该行人框中对角线上的两个点是否都在待检测四边形区域内，如果同时在四边形区域内，则对该人员进行行人检测。如图 5-5 所示。

图 5-5 基于区域入侵的行人检测

5.2.2 基于 OSFANet 网络的行为识别

非法过闸机行为识别的研究我们在前面两章已经进行了实验分析，最终选择使用添加注意力机制的 OSFANet 网络进行非法过闸机的行为识别。针对经过视频流切分后的视频序列，我们对单帧图片进行依次加载，然后读取划分待检测区域四边形的四个坐标，对闸机口待检测区域使用 YOLOv5 算法检测闸机口附近的行人，对检测到的行人我们用 AlphaPose 骨架提取算法对每个行人生成对应的骨架图，然后使用三个 list（列表）对行人框，原图，骨架图进行保存。然后对 list 进行遍历，把原图和骨架图输入到 OSFANet 网络，实现非法过闸机的行为识

别。行为识别的整体框图如图 5-6 所示。

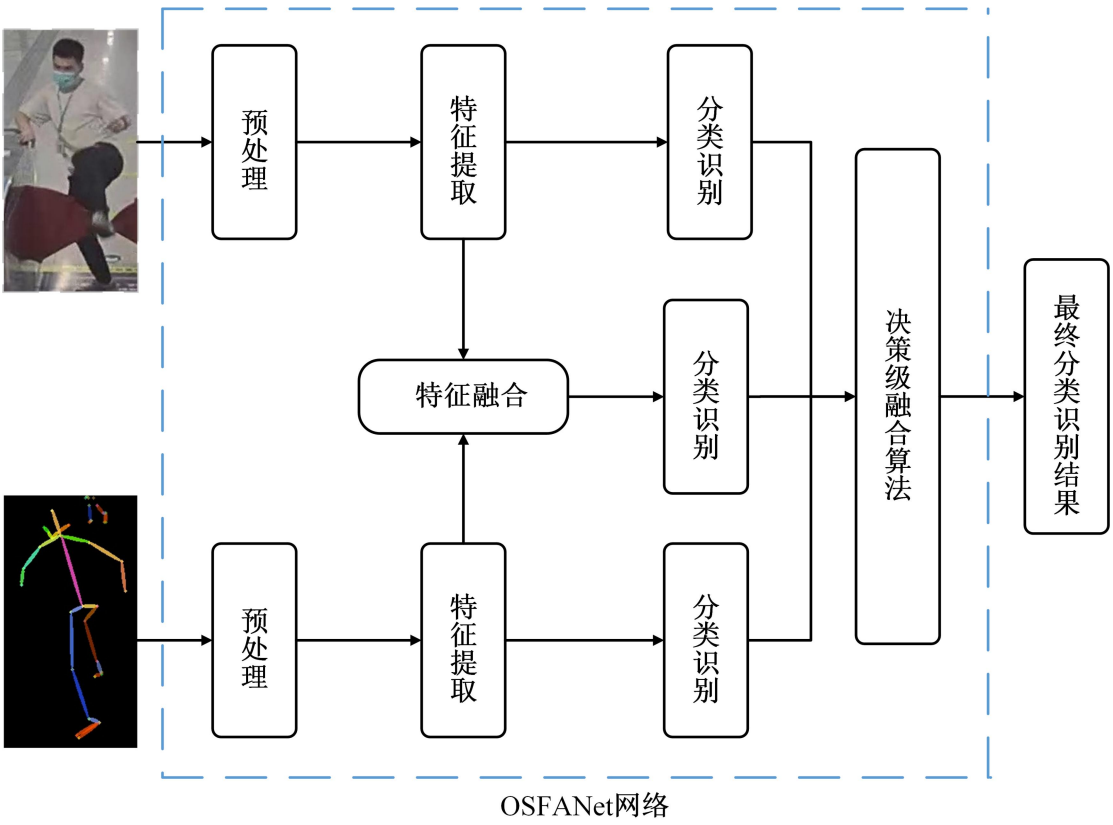


图 5-6 行为识别网络整体框图

5.2.3 基于滑动窗口的行人跟踪设计

5.2.3.1 IOU 介绍

IOU 又称交并比，如图 5-7 所示，是一种简单的测量标准，这个标准主要是用来比较两个物体之间的相关度，简单地说就是两个区域之间的重合度。重合度越高，IOU 的值就越大，重合度越低，IOU 的值也就越小。IOU 的计算公式为两个区域重叠的部分除以它们的集合，也就是两个区域的交叉面积除以两个区域的合并面积。最后得到的结果我们称为 IOU 的值。

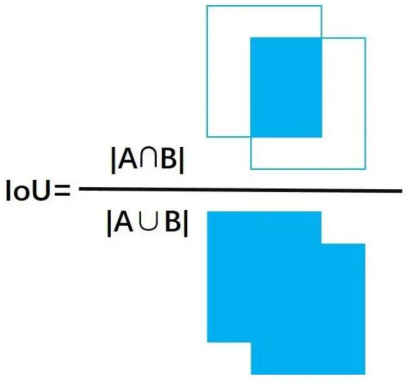


图 5-7 交并比示意图

5.2.3.2 跟踪算法设计

对于视频流中的过闸机人员，我们用 YOLOv5 算法进行行人检测，用 AlphaPose 提取对应的骨架，用 OSFANet 网络进行非法行为预测，最后我们会在每帧图片中定义一个列表，把图中出现过闸机的人员信息以对象的形式保存在列表中来实现跟踪[88]。其中每个对象的信息包括：人员 ID，行人框的坐标，是否存在非法过闸机行为的 flag。对上一帧图片中的人和下一帧图片中的人员以遍历的方式进行 IOU 匹配，如果 IOU 值大于 0.95 则判断上一张图片的这个人和下一张图片的这个人是同一个人，对同一个人中的每帧非法行为结果存入到一个长度为 10 的队列中， 该队列又称为视频流中的滑动窗口，如图 5-8 所示。

图 5-8 基于滑动窗口的行人跟踪设计

随视频流进行滑动的同时进行非法行为判断。我们规定发生非法行为的人员 flag 置为 1，未发生非法过闸机行为的 flag 置为 0，我们设定如果该人员的 flag 信息在滑动窗口的 10 帧中有 8 帧或 8 帧以上 flag 都是 1，也就是 10 帧中出现 8 帧或 8 帧以上出现了非法过闸机行为，则判断该人员正在进行非法过闸机行为，我们会对该人员进行报警，把该人员用红色框框出的同时对框出的部分以图片的形式保存到报警文件夹。对于没有非法行为的人员，我们用蓝色的框进行框出如图 5-9 所示。

5.3 基于视频事件的非法过闸机行为检测算法实现

上文已经介绍了非法过闸机行为检测的算法设计，为了验证算法的有效性，海信公司的测试人员召集了公司的员工在地铁闸机口进行了模拟乘客过闸机的行为，并专门录制了 10 段视频作为考核标准来验证非法过闸机行为检测算法的有效性，其中第一段视频中有 58 个非法过闸机行为，73 个正常过闸机行为。第二段视频中有 60 个非法过闸机行为，81 个正常过闸机行为。第三个视频中有 42 个非法过闸机行为，56 个正常过闸机行为。第四段视频中有 78 个非法过闸机行为，103 个正常过闸机行为。第五段视频中有 36 个非法过闸机行为，54 个正常过闸机行为。第六段视频中有 66 个非法过闸机行为，94 个正常过闸机行为。第七段视频中有 47 个非法过闸机行为，76 个正常过闸机行为。第八段视频中有 63 个非法过闸机行为，98 个正常过闸机行为。第九段视频中有 49 个非法过闸机行为，87 个正常过闸机行为。第十段视频中有 65 个非法过闸机行为，107 个正常过闸机行为。这 10 段视频总共有 564 个非法过闸机行为，我们用这 10 段视频来验证非法过闸机行为检测算法。实验环境如下表 5-1 所示。

表 5-1 实验环境

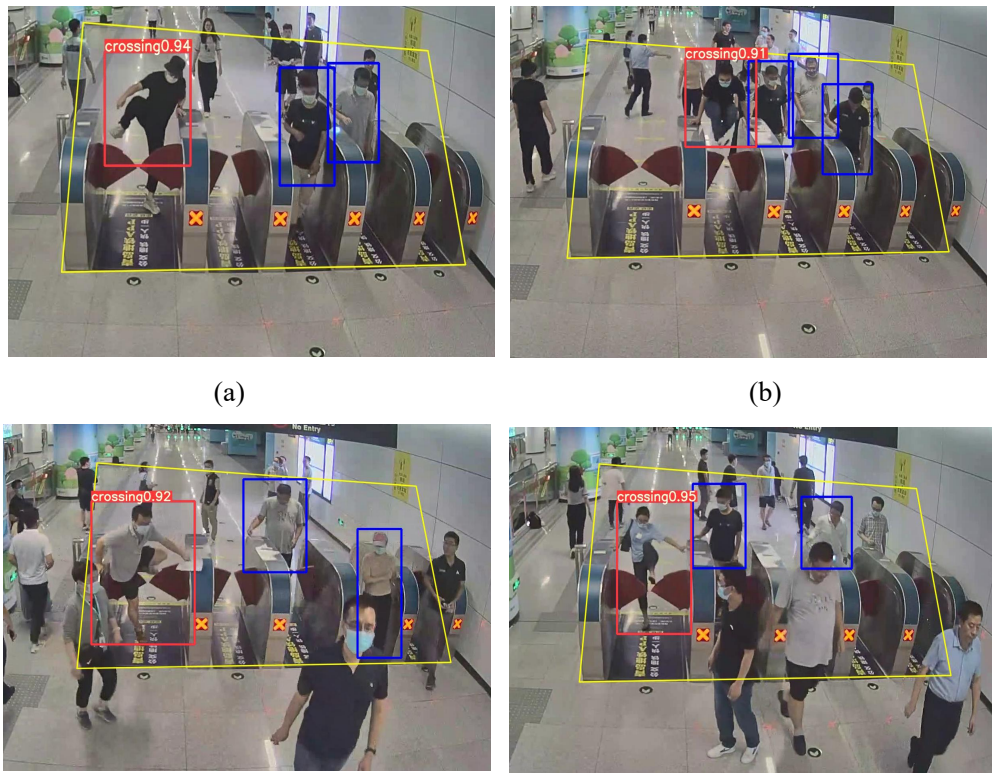
软硬件环境	型号
CPU	Intel Xeon Gold 6240
显卡	Tesla V100 SXM2
显存	32G
内存	1TB
Python 版本	3.8
Pytorch 版本	1.6
CUDA 版本	10.2

我们定义一个人完整地经过闸机这一过程称为一个事件，测试是基于事件的非法过闸机行为作为判断指标。视频中有非法过闸机行为事件，也有正常过闸机行为事件，而测试主要是针对非法过闸机行为事件进行统计，视频要求以 1 秒 10 帧的速度进行测试，也就是每帧的平均处理时间不能大于 100ms，而我们设计的非法过闸机行为检测算法处理每帧图片的平均时间大约为 90ms，处理速度能够满足基本要求，所设计的滑动窗口的大小是 10 帧，也就是 1 秒的检测时间，在 1 秒内，10 帧图片有 8 帧以上出现非法过闸机行为我们则判断该人员发生非法过闸机行为，对该人员进行报警。我们把算法正确地检测到非法过闸机行为事件的个数称为正检数。把正常过闸机的行为事件误检为非法过闸机行为事件的个数称为误检数。把非法过闸机行为事件未检测出来的个数称为漏检数。我们定义正检率为视频中正检数占正检数和误检数之和的百分比，也就是精确率。召回率为正检数占正检数和漏检数之和的百分比。测试结果如表 5-2 所示。

表 5-2 非法过闸机行为检测事件测试结果

视频	非法事件总数	正检数	误检数	漏检数	正检率	召回率
视频 1	58	51	4	7	92.73%	87.93%
视频 2	60	53	5	7	91.38%	88.33%
视频 3	42	37	2	5	94.87%	88.10%
视频 4	78	71	5	7	93.42%	91.03%
视频 5	36	32	3	4	91.43%	88.89%
视频 6	66	58	6	8	90.63%	87.88%
视频 7	47	42	3	5	93.33%	89.36%
视频 8	63	57	5	6	91.94%	90.48%
视频 9	49	44	4	5	91.67%	89.80%
视频 10	65	58	5	7	92.06%	89.23%
总计	**564**	**503**	**42**	**61**	**92.29%**	**89.18%**

由表可知，本文的非法过闸机行为检测算法在测试的三段视频中对非法过闸机事件的正检率在 92%左右，召回率在 89%左右。整体来说，多帧的综合判断结果优于单帧的结果。在实际工程中，一般都会尽量避免误检的情况发生，而可以允许少量漏检的情况，为了平衡误报和漏报的情况，本文设计的滑动窗口阈值设置为 0.8，也就是说 10 帧图片需要满足 8 帧以及 8 帧以上才能判断为非法过闸机行为，由于阈值设置较高，并且有发生漏检的情况，从而导致召回率提升没有精确率多。基于深度学习的非法过闸机行为检测算法的检测效果如图 5-9 所示。



(a)　(b)

(c)　(d)

图 5-9 基于视频事件的非法过闸机行为检测算法实现

从图 5-9 中可知，本文提出的非法过闸机行为检测算法在视频中的检测效果还是可以的，在大多数情况下都可以检测出来，有时候偶尔会出现在区域内检测不到人的情况，主要是地铁内浅色墙壁以及地砖颜色有关，这种浅色背景会导致浅色衣服的人出现漏检情况。还有一些行人遮挡情况也是不好检测出来的。整体来说本文提出的非法过闸机行为检测算法效果是可以的，针对跨越闸机和翻越闸机都有很好的检测效果，目前正处于海信公司试运行阶段。

为了验证 OSFANet 网络在基于视频的非法过闸机行为检测上的有效性，本文使用添加 CBAM 注意力机制的 ResNet50 网络和中期融合网络进行基于视频的非法过闸机行为检测实验。同样使用上文中 10 段视频的 564 个非法过闸机行为事件作验证，实验对比结果如表 5-3 所示。

表 5-3 不同算法基于视频的非法过闸机行为检测实验结果对比

模型	正检数	误检数	漏检数	正检率	召回率	FPS
ResNet50+CBAM	476	80	88	85.61%	84.40%	**17**
MiddleMergeModel	488	56	76	89.71%	86.52%	9
OSFANet	**503**	**42**	**61**	**92.29%**	**89.18%**	11

从表 5-3 中可知，OSFANet 算法在基于视频的非法过闸机行为检测实验中取得的效果最好，由于本文设计的非法过闸机行为检测算法需要先检测再分类，基于信息融合的方法还需要骨架提取，因此 FPS 都不是很高，虽然不能满足安防监控视频摄像头每秒 25 帧的实时性，但是也是可以基本满足海信公司的测试要求。在满足测试要求的情况下，本文提出的 OSFANet 算法在正检率和召回率上都是最优的，实验结果很好地证明 OSFANet 网络在基于视频的非法过闸机行为检测上的有效性。

5.4 本章小结

本章主要是介绍基于视频的非法过闸机行为检测算法的设计与实现，整体来说可以分为三个部分，先对视频流进行切分，使用基于区域入侵的 YOLOv5 目标检测算法对每帧图片进行行人检测，并使用 AlphaPose 骨架提取算法获得检测到的行人图片对应的骨架信息，然后使用 OSFANet 网络对原图和骨架图进行行为识别，获得单帧的非法过闸机行为识别结果。最后采用基于滑动窗口的跟踪算法获得多帧图片的预测结果来进行行为分析，最终判断该人员是否存在非法过闸机行为。最后采用基于视频事件的方式对 564 个非法过闸机行为事件进行非法过闸机行为检测的测试。经过测试，该算法的平均正检率为 92.29%，平均召回率为 89.18%。随后，本文对比了添加注意力机制 CBAM 的 ResNet50 网络，以及中期融合网络进行基于视频的非法过闸机行为检测实验。结果表明本文提出的 OSFANet 网络在非法过闸机行为检测的实验中在检测效果更好。

第 6 章 总结与展望

6.1 总结

闸机口是一种常见的公共防范措施和安全进出入设施，常见于地铁进出刷卡、小区门禁刷卡已经重大安检地方，其能够在物理上实现对空间的隔离与人员进出入的阻拦。而闸机口本身是由两侧的分流栏杆和中间的挡板组成，挡板作为对人的阻拦，更大程度上在告示要在获得允许后才能进出入。在一般情况下，并不能实现对人的阻拦的，即人员可以违规违法地轻易采用其他措施（翻越、跨越等）进行通过，在地铁场景下，这些行为统称为非法过闸机。进出入地铁应正确性刷卡或者投币，但是违规的通过行为会给地铁管理带来财产上损失，因此必须给这些有非法行为的人员进行检测和报警，然而地铁闸机口的广泛，采用人为的看护是极其耗时耗力的。近年来，随着计算机视觉技术和人工智能的飞速发展，对图像和视频的处理能力的愈发增强，以监控视频为基础的行为分析也取得了突飞猛进的发展，因此，通过视频分析技术判断过闸机的行为的也具有一定的可行性。

近年来，基于单张图片的方法进行行人行为分析已广泛应用于现实生活中，给人们的生产生活带来了极大的便捷。其中，目标检测算法已经广泛用于生活场景图像检测与识别；骨架检测技术也同样开始广泛应用于人的行为和姿态的分析识别。但是，当前的目标检测和骨架检测技术大都是基于单帧图片的，无法从视频中的一系列连续动作判断人员的行为，直接应用于非法过闸机的行为检测是不可行的。因此结合时间维度的三维信息来综合评定过闸机的人员是否存在非法过闸机行为事件有着极大的应用前景。

本文提出了一种基于深度学习的非法过闸机行为检测算法，目的是检测地铁闸机口的人员是否存在非法过闸机行为，并对发生非法过闸机行为的人员进行相应的行为告警。主要的技术路线是通过将地铁的监控视频流进行截取获取连续的图像序列，对单帧图片划分待检测区域的四边形；引入 YOLOv5 目标检测算法和 AlphaPose 骨架提取算法，提取待检测区域中的人员信息和骨架信息，从而构建人员信息；再基于人员信息构建行为识别网络，对人员行为进行识别，判断是否发生违法行为；最后基于单帧的预测结果对连续多帧图片使用相应的跟踪算法和逻辑规则判断视频中的人员是否存在非法过闸机行为，对检测的违法行为进行合理报警。本研究能够对非法过闸机的行为进行很好的识别，大大提升了地铁周边区域的安全性和合法性。本文的具体研究工作可以分为以下几个部分：

（1）本文制作了一个基于地铁闸机口场景的非法过闸机行为识别数据集。数据集的建立是研究非法过闸机行为检测的基础，本文对海信公司提供的地铁闸

机口行人过闸机的视频进行视频抽帧处理，将视频流切分为图像序列，通过YOLOv5 的目标检测算法对待检测区域的行人进行检测，对图像中闸机口检测框内的行人进行保存，形成基于原始图像的非法过闸机行为识别数据集。然后使用AlphaPose 骨架提取技术对检测框的图像进行骨架提取，构建了基于骨架的非法过闸机行为识别数据集。

（2）本文对非法过闸机行为识别数据集进行了基于原图和骨架图分类研究，在原图和骨架图分类实验中，本文选取了不同的分类网络，分别对原始图像数据集和骨架数据集进行训练。最后使用海信公司提供的测试视频进行视频切分为图片序列，对测试图片进行标注后使用训练好网络进行预测，通过对比实验得知ResNet50 分类效果更好，随后对 ResNet50 网络进行添加注意力机制的实验，发现添加 CBAM 模块的 ResNet50 分类效果提升较多。紧接着进行了原图信息和骨架信息相结合的对比实验，由实验可知中期特征融合的分类效果更好。最后本文提出了一种添加注意力机制的 OSFANet 网络，相较于单种数据的预测网络和中期融合网络有更好的分类效果，并做了消融实验验证 OSFANet 的有效性。

（3）本文设计了一种基于深度学习的非法过闸机行为检测算法，对闸机口待检测区域内使用区域入侵算法进行行人框对角点是否在区域内的判断，对区域内的行人进行行人检测，对检测到的行人进行对应的骨架提取，构建行人行为信息，使用 OSFANet 网络进行非法过闸机行为预测，使用基于滑动窗口的方法对多帧图片中的人员进行跟踪并结果多帧的预测结果判断该人员是否存在非法过闸机行为，对正常过闸机的人员使用蓝色框框出即可，对发生非法行为的人员用红色框框出并保存为图片上传至对应的报警文件夹。

6.2 展望

本文提出的基于深度学习的地铁非法过闸机行为检测算法可以实现地铁闸机口行人过闸机时的非法行为检测和报警，目前正处于海信公司试运行测试阶段。目前为止该算法还存在着可改进的空间：

（1）针对非法过闸机的行为识别效果并不能达到 100%的准确，我们需要训练更多的不同非法过闸机行为数据来提升该算法的泛化能力。

（2）针对非法过闸机行为检测算法模型内存较大问题，后续我们需要进行模型轻量化部署。

参考文献

[1] 骆云志,刘治红.视频监控技术发展综述[J].兵工自动化,2009.

[2] 万家兵,霍宏.智能视频监控中目标的检测与识别[M].上海交通大学出版社,2010.

[3] 徐杨,吴成东,陈东岳.基于视频图像的交通事件自动检测算法综述[J].计算机应用研究, 2011,28(4): 1206-1210

[4] 王龙.基于视频分析的客流检测子系统的设计与实现[D].北京交通大学，2012.

[5] 马天文.地铁安全运营的影响因素与对策探究[J].科技创新与应用, 2016(32).

[6] 杨静,韩丽东.人工智能双目视觉闸机的可靠性分析[J]. 机械工业标准化与质量, 2020(02):41-45.

[7] 吴帆.地铁应用场景中群体异常行为检测研究[D]. 东南大学 , 2019.

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

[9] Viola P, Jones M. Rapid object detection using a boosted cascade of simple features[C]//Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001. Ieee, 2001, 1: I-I.

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

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

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

[13] Mohan A, Papageorgiou C, Poggio T. Example-based object detection in images by components[J]. IEEE transactions on pattern analysis and machine intelligence, 2001, 23(4): 349-361.

[14] Ke Y, Sukthankar R. PCA-SIFT: A more distinctive representation for local image descriptors[C]//Proceedings of the 2004 IEEE Computer Society

Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004. IEEE, 2004, 2: II-II.

[15] Felzenszwalb P, McAllester D, Ramanan D. A discriminatively trained, multiscale, deformable part model[C]//2008 IEEE conference on computer vision and pattern recognition. Ieee, 2008: 1-8.

[16] Felzenszwalb P F, Girshick R B, McAllester D, et al. Object detection with discriminatively trained part-based models[J]. IEEE transactions on pattern analysis and machine intelligence, 2009, 32(9): 1627-1645.

[17] Felzenszwalb P F, Girshick R B, McAllester D. Cascade object detection with deformable part models[C]//2010 IEEE Computer society conference on computer vision and pattern recognition. Ieee, 2010: 2241-2248.

[18] 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. 2014: 580-587.

[19] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[J]. Advances in neural information processing systems, 2012, 25.

[20] He K, Zhang X, Ren S, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition[J]. IEEE transactions on pattern analysis and machine intelligence, 2015, 37(9): 1904-1916.

[21] Grauman K, Darrell T. The pyramid match kernel: Discriminative classification with sets of image features[C]//Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1. IEEE, 2005, 2: 1458-1465.

[22] Zeiler M D, Fergus R. Visualizing and understanding convolutional networks[C]//European conference on computer vision. Springer, Cham, 2014: 818-833.

[23] Girshick R. Fast r-cnn[C]//Proceedings of the IEEE international conference on computer vision. 2015: 1440-1448.

[24] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[J]. Advances in neural information processing systems, 2012, 25.

[25] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv:1409.1556, 2014.

[26] Vedaldi A, Jia Y, Shelhamer E, et al. Convolutional architecture for fast feature

embedding[J]. Cornell University, 2014.

[27] Ren S, He K, Girshick R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks[J]. Advances in neural information processing systems, 2015, 28.

[28] 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.

[29] Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 1-9.

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

[31] Liu W, Anguelov D, Erhan D, et al. Ssd: Single shot multibox detector[C]//European conference on computer vision. Springer, Cham, 2016: 21-37.

[32] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv:1409.1556, 2014.

[33] Erhan D, Szegedy C, Toshev A, et al. Scalable object detection using deep neural networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2014: 2147-2154.

[34] Felzenszwalb P, McAllester D, Ramanan D. A discriminatively trained, multiscale, deformable part model[C]//2008 IEEE conference on computer vision and pattern recognition. Ieee, 2008: 1-8.

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

[36] Redmon J. Darknet: Open source neural networks in c[J]. 2013.

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

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

[39] Bochkovskiy A, Wang C Y, Liao H Y M. Yolov4: Optimal speed and accuracy of object detection[J]. arXiv preprint arXiv:2004.10934, 2020.

[40] Zheng Z, Wang P, Liu W, et al. Distance-IoU loss: Faster and better learning for bounding box regression[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2020, 34(07): 12993-13000.

[41] Loshchilov I, Hutter F. Sgdr: Stochastic gradient descent with warm restarts[J]. arXiv preprint arXiv:1608.03983, 2016.

[42] Liu S, Qi L, Qin H, et al. Path aggregation network for instance segmentation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 8759-8768.

[43] Jocher G, Stoken A, Chaurasia A, et al. Ultralytics/yolov5: v6. 0—YOLOv5n "Nano" Models, Roboflow Integration, Tensor, Flow Export, Open, CV DNN Support[J]. Zenodo Tech. Rep., 2021.

[44] Candamo J, Shreve M, Goldgof D B, et al. Understanding transit scenes: A survey on human behavior-recognition algorithms[J]. IEEE transactions on intelligent transportation systems, 2009, 11(1): 206-224.

[45] Zhang Z, Tao D. Slow feature analysis for human action recognition[J]. IEEE transactions on pattern analysis and machine intelligence, 2012, 34(3): 436-450.

[46] Thuan D. Evolution of yolo algorithm and yolov5: the state-of-the-art object detection algorithm[J]. 2021.

[47] Herath S, Harandi M, Porikli F. Going deeper into action recognition: A survey[J]. Image and vision computing, 2017, 60: 4-21.

[48] Das Dawn D, Shaikh S H. A comprehensive survey of human action recognition with spatio-temporal interest point (STIP) detector[J]. The Visual Computer, 2016, 32(3): 289-306.

[49] Laptev I. On space-time interest points[J]. International journal of computer vision, 2005, 64(2): 107-123.

[50] Dollár P, Rabaud V, Cottrell G, et al. Behavior recognition via sparse spatio-temporal features[C]//2005 IEEE international workshop on visual surveillance and performance evaluation of tracking and surveillance. IEEE, 2005: 65-72.

[51] Hu J F, Wang X H, Zheng W S, et al. Research progress and prospect of RGB-D behavior recognition[J]. J Autom, 2019, 45(5): 829-840.

[52] Khaire P, Kumar P, Imran J. Combining CNN streams of RGB-D and skeletal data for human activity recognition[J]. Pattern Recognition Letters, 2018, 115: 107-116.

[53] Franco A, Magnani A, Maio D. A multimodal approach for human activity recognition based on skeleton and RGB data[J]. Pattern Recognition Letters, 2020, 131: 293-299.

[54] Scovanner P, Ali S, Shah M. A 3-dimensional sift descriptor and its application to action recognition[C]//Proceedings of the 15th ACM international conference on Multimedia. 2007: 357-360.

[55] Yilmaz A, Shah M. Actions sketch: A novel action representation[C]//2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05). IEEE, 2005, 1: 984-989.

[56] Klaser A, Marszałek M, Schmid C. A spatio-temporal descriptor based on 3d-gradients[C]//BMVC 2008-19th British Machine Vision Conference. British Machine Vision Association, 2008: 275: 1-10.

[57] Wang H, Kläser A, Schmid C, et al. Dense trajectories and motion boundary descriptors for action recognition[J]. International journal of computer vision, 2013, 103(1): 60-79.

[58] Wang J, Xia L. Abnormal behavior detection in videos using deep learning[J]. Cluster Computing, 2019, 22(4): 9229-9239.

[59] Jaouedi N, Boujnah N, Bouhlel M S. A new hybrid deep learning model for human action recognition[J]. Journal of King Saud University-Computer and Information Sciences, 2020, 32(4): 447-453.

[60] Harris C, Stephens M. A combined corner and edge detector[C]//Alvey vision conference. 1988, 15(50): 10-5244.

[61] Willems G, Tuytelaars T, Gool L V. An efficient dense and scale-invariant spatio-temporal interest point detector[C]//European conference on computer vision. Springer, Berlin, Heidelberg, 2008: 650-663.

[62] Gorelick L, Blank M, Shechtman E, et al. Actions as space-time shapes[J]. IEEE transactions on pattern analysis and machine intelligence, 2007, 29(12): 2247-2253.

[63] Sahoo S P, Srinivasu U, Ari S. 3D Features for human action recognition with semi-supervised learning[J]. IET Image Processing, 2019, 13(6): 983-990.

[64] Huang Q Q, Zhou F Y, Liu M Z. Survey of human action recognition algorithms based on video[J]. Application Research of Computers, 2020, 37(11): 19-25.

[65] 朱旭东. 基于语义主题模型的人体异常行为识别研究[D].西安电子科技大

学,2011.

[66] Yang X, Zhang C, Tian Y L. Recognizing actions using depth motion maps-based histograms of oriented gradients[C]//Proceedings of the 20th ACM international conference on Multimedia. 2012: 1057-1060.

[67] Chen C, Liu K, Kehtarnavaz N. Real-time human action recognition based on depth motion maps[J]. Journal of real-time image processing, 2016, 12(1): 155-163.

[68] Wang H, Schmid C. Action recognition with improved trajectories[C]// Proceedings of the IEEE international conference on computer vision. 2013: 3551-3558.

[69] Yi Y, Zheng Z, Lin M. Realistic action recognition with salient foreground trajectories[J]. Expert Systems with Applications, 2017, 75: 44-55.

[70] Abdul-Azim H A, Hemayed E E. Human action recognition using trajectory-based representation[J]. Egyptian Informatics Journal, 2015, 16(2): 187-198.

[71] Ji S, Xu W, Yang M, et al. 3D convolutional neural networks for human action recognition[J]. IEEE transactions on pattern analysis and machine intelligence, 2012, 35(1): 221-231.

[72] Tran D, Bourdev L, Fergus R, et al. Learning spatiotemporal features with 3d convolutional networks[C]//Proceedings of the IEEE international conference on computer vision. 2015: 4489-4497.

[73] Tran D, Ray J, Shou Z, et al. Convnet architecture search for spatiotemporal feature learning[J]. arXiv preprint arXiv:1708.05038, 2017.

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

[75] Qiu Z, Yao T, Mei T. Learning spatio-temporal representation with pseudo-3d residual networks[C]//proceedings of the IEEE International Conference on Computer Vision. 2017: 5533-5541.

[76] Simonyan K, Zisserman A. Two-stream convolutional networks for action recognition in videos[J]. Advances in neural information processing systems, 2014, 27.

[77] Zhao Y, Man K L, Smith J, et al. Improved two-stream model for human action recognition[J]. EURASIP Journal on Image and Video Processing, 2020, 2020(1):

1-9.

[78] Li C, Zhong Q, Xie D, et al. Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation[J]. arXiv preprint arXiv:1804.06055, 2018.

[79] Tang Y, Tian Y, Lu J, et al. Deep progressive reinforcement learning for skeleton-based action recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 5323-5332.

[80] Rahmani H, Bennamoun M. Learning action recognition model from depth and skeleton videos[C]//Proceedings of the IEEE International Conference on Computer Vision. 2017: 5832-5841.

[81] Khaire P, Kumar P, Imran J. Combining CNN streams of RGB-D and skeletal data for human activity recognition[J]. Pattern Recognition Letters, 2018, 115: 107-116.

[82] Fang H S, Xie S, Tai Y W, et al. Rmpe: Regional multi-person pose estimation[C]//Proceedings of the IEEE international conference on computer vision. 2017: 2334-2343.

[83] Cao Z, Simon T, Wei S E, et al. Realtime multi-person 2d pose estimation using part affinity fields[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 7291-7299.

[84] Hu J, Shen L, Albanie S, et al. Squeeze-and-Excitation Networks.(May 2019)[J]. 2019.

[85] Park J, Woo S, Lee J Y, et al. Bam: Bottleneck attention module[J]. arXiv preprint arXiv:1807.06514, 2018.

[86] Woo S, Park J, Lee J Y, et al. Cbam: Convolutional block attention module[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 3-19.

[87] 向翼凌,何伟. 面向校园安全的视频区域入侵检测算法. 武汉工程大学学报,2019, 41(1): 93–97.

[88] 徐涛,马克,刘才华.基于深度学习的行人多目标跟踪方法[J].吉林大学学报(工学版), 2021, 51(01):27-38.

攻读学位期间主要研究成果

一、已授权发明专利与软件著作权

[1] 谢斌，万思远．一种 ki67 指数确定方法及系统．中国：202110353954 .0 [P]，2021-04-01．已授权．

[2] 中南大学．基于异构信息的胶质瘤分子分型预测系统 V1.0，软著申请登记号：2021SR0693511，证书号：软著登记第 7416137 号，开发完成时间：2021 年 02 月 10 日，申请时间：2021 年 05 月 14 日（本人第一发明人）．

二、发表学术论文

[1] Xuhui Li,Yong Xu,Feng Xiang,Siyuan Wan,Weihong Huang,Bin Xie. Prediction of IDH Mutation Status of Glioma Based on Multimodal MRI Images[C]//2021 3rd International Conference on Intelligent Medicine and Image Processing. 2021: 39-44.. https://doi.org/10.1145/3468945.3468952.

三、参与科研项目

（1） 企业合作研究课题：基于机器视觉的地铁内非法行为检测算法设计和系统开发，2021.1-至今，主要研究人员．

（2） 教育部-中国移动医疗联合实验室开放课题：“基于非侵入式信息的胶质瘤分子分型预测系统及临床应用评估”，2019.7- 2020.11，主要研究人员．

（3） 企业合作研究课题：书法智能评测系统，2020.4-2021.5,主要开发人员．

四、硕士期间参加学术交流活动

（1） 2020 年 11 月 “之识无界”大会暨第三届之江国际青年人才论坛（杭州）．

五、硕士期间获得奖励情况

（1） 省部级，第一届湖南省研究生人工智能创新大赛一等奖，2020.

（2） 校级，中南大学一等学业奖学金，2021.

致 谢

光阴似箭，岁月如梭，转眼间三年的研究生学习生涯即将画上句号。我十分有幸能就读于中南大学，在这三年间，是我成长最快的阶段。在这里，我享受着学校给予的优质教育教学资源，结交了许多优秀的良师益友，学习到了前沿的知识，了解了广阔的社会，有过艰苦的熬夜学习生活，也经历过许多美好的时光。在此临近毕业之际，谨向对所有给予我关心、支持和帮助的各位老师、同学以及家人表示最诚挚的感谢与最美好的祝愿。

首先感谢我的导师谢斌教授，谢老师是一位知识渊博的老师，第一次和谢老师见面还是在夏令营的时候，谢老师对人工智能领域的见解以及对科研敏锐的洞察力和前瞻性开阔了我的眼界，提高了我的格局，让我心生崇拜，从那一刻起我就立志要成为谢老师的学生。功夫不负有心人，经过面试层层选拔，我终于加入了谢老师的人工智能与机器人实验室。从此，开始了我研究生阶段的科研生涯之旅。加入智能所 AIRL 人工智能与机器人实验室，让我学到了优秀的师兄师姐身上的努力学习精神，在谢斌老师的指导下展开自己的科研学习，实验室每周的例会格外的期待，因为在例会上可以见证每个人一周的学习动态，见贤思齐，针对一些比较棘手的难题，我们会在一起讨论，可以跟实验室勤奋优秀的同学们一起相互学习共同进步。谢老师不仅可以在科研上给予我帮助，还教会了我许多人生哲理，是我人生道路上的指明灯。从某种意义上来讲，谢老师不仅是一位科研学者，还是思想家，哲学家。除此之外，还要感谢智能所的蔡自兴老师，王勇老师，陈白帆老师，王冰川老师，邹逸群老师，雷琪老师，李仪老师，刘丽珏老师等，感谢你们的精心授课和在学业上给予的指导！

感谢 AIRL 实验室的师兄师姐师弟师妹们！感谢李沁师姐、何小宇师兄、宋迪师兄、张文婷师姐、秦觅师兄、林学斌师兄、徐勇师兄、赵新宇师兄、丘文杰师兄、程清师兄、彭哲师兄做的好榜样，感谢吴迪、孙铭蔚、李旭辉、后国炜、邝逸灵、解振南、王清锋、聂海涛、宫心玉、赵嘉明实验室的小伙伴，让我的读研生活丰富多彩。你们对学术的专注和努力学习的精神让我敬佩，是你们在我遇到问题时，能够不厌其烦地帮助我，让我得以成长！特别感谢徐勇师兄和李旭辉师弟在科研上的帮助，感谢程清师兄和彭哲师兄在我找实习时对我的帮助。感谢靳振涛，钟立科，王晓晶，杨俊，张鹏飞等 209 的小伙伴，感谢你们陪我一起攻克科研难题，互相鼓励，共同进步。感谢已经毕业但仍然为我们答疑解难的前辈们，再次感谢智能所师兄师姐师弟师妹们在这三年来对我的帮助！

感谢我的舍友靳振涛、罗海森、华成尧和隔壁舍友钟立科、齐帅宾、汤超宇！

谢谢你们陪我度过了难忘的研究生生活和一直以来对我的关心和照顾，谢谢你们带着我一起逛街、爬山、撸串、探索长沙的美好。

最后，深深的感谢我的家人们，你们含辛茹苦把我养大，不计一切在求学路上始终默默地支持着我供我读书。是你们不求回报的奉献才让我取得了今天的成果，言树之背，养育之恩，我唯有通过不懈努力的奋斗才能报答此恩情之一二。

2021 年 5 月

万思远