目录

[ГЛАВА 1 Обзор литературы （文件综述） 1](#_Toc134752515)

[目标检测 1](#_Toc134752516)

[基于传统方法的目标检测 2](#_Toc134752517)

[基于深度学习的目标检测 2](#_Toc134752518)

[YOLOv5目标检测算法 3](#_Toc134752519)

[多目标跟踪 4](#_Toc134752520)

[基于目标检测的多目标跟踪 (Tracking by Detection) 5](#_Toc134752521)

[联合检测和跟踪的多目标跟踪（Join Detection and Tracking, JDT） 5](#_Toc134752522)

[DeepSORT目标跟踪算法 6](#_Toc134752523)

[行人重识别 8](#_Toc134752524)

[基于表征学习(Representation learning)的方法 9](#_Toc134752525)

[基于度量学习 (Metric learning)的方法 11](#_Toc134752526)

[Fast ReID算法 12](#_Toc134752527)

[人流量统计 13](#_Toc134752528)

[本文的主要研究内容 14](#_Toc134752529)

# ГЛАВА 1 Обзор литературы （文件综述）

人流量统计是一个集目标检测、多目标跟踪、行人重识别技术于一体的综合性任务。鉴于此，本文首先叙述目标检测和多目标跟踪的研究现状~~，其次考察跨摄像头多目标跟踪的研究现状~~。考虑到行人重识别是解决跨摄像头轨迹关联问题的核心技术，在本文最后也总结了当前主流的行人重识别方法以及现存的一些不足，借此引入行人重识别的新方法-FastReID。

## 目标检测

目标检测是多目标跟踪任务的基础，其主要目的是把需要被检测的目标从每一帧的静态图片中用矩形框标记出来，得到目标在图片中的具体位置。如果目标检测中出现了漏检，那么后期也就不会对目标进行跟踪，所以目标检测在多目标跟踪任务中起着非常关键的作用。纵观目标检测的整个发展过程，可将其分为基于传统方法的目标检测和基于深度学习的目标检测。

狗站在自行车后面

描述已自动生成

### 基于传统方法的目标检测

在深度学习发展之前，传统的目标检测算法通常遵循基于手工特征设计的思想，此类方法大致包括三个步骤，分别为区域选择、特征提取和分类器分类。该时期主流的检测算法包括 Viola-Jone 检测器[2]、HOG 检测器[3]和变形部件模型[4]（Deformable Parts Model, DPM）。Viola 等人提出的 Viola-Jone 检测器采用积分图像、特制选择和检测级联三种有效的策略，并通过从左到右、从上到下的滑动方式，判断是否有窗口包含目标，该检测器的提出使得目标检测的速度大大提高。HOG 检测器最初由 Dalal 等人在 2005年提出，其核心思想是局部物体形状和外观能够被局部梯度或边缘的密度分布所描述，利用方向梯度直方图特征描述子来构建特征，并与SVM分类器结合，有效地解决了行人检测任务。另外，DPM 检测器由 Pedro 等人在 2008 年提出，遵循“分而治之”的理论对目标进行检测， 通过检测目标包含的各个部件从而完成对目标整体的检测。

### 基于深度学习的目标检测

随着传统目标检测算法的性能趋于饱和，目标检测的研究也进入了一个平稳的发展期，直到卷积神经网络的出现，使得目标检测技术的性能获得了显著的提高。此类算法大致可分为以下两类，一类是两阶段检测算法，先在图像中选择候选区域，然后再完成这些区域的分类任务，该类算法的特点是精度高但速度慢。如 Ross 等人[5]所提出的 R-CNN，首先以选择性搜索操作预先定位图片中的待检测区域，然后通过卷积神经网络提取这些区域上的特征，最后进行分类。虽然 R- CNN 算法大大提高了目标检测精度，但其存在计算冗余的问题。He 等人[6]提出的 SPPNet 算法，通过加入空间金字塔池化层解决了 R-CNN 反复缩放图像造成的影响。此外，Girshick 等人[7]提出 Fast R-CNN 算法，通过加入区域池化层进一步优化了 R-CNN 与 SPPNet 算法。Ren 等人[8]提出 Faster R-CNN 算法，将选择性搜索用区域提案网络代替，对网络进行端到端地训练，极大地提高了检测器的速度。另一类是一阶段检测算法，在产生候选区域的同时进行回归输出检测对象的位置和类别，其代表性算法就是 YOLO 系列算法[9]和 SSD 算法[10]。YOLO 系列算法首先由 Redmon 等人于 2016 年提出，其抛弃了“提案检测+验证”的两阶段检测范式，通过将图像分割成多个网格，直接预测每个网格所属的类别概率和位置信息。针对 YOLO 算法检测精度低的缺点，Liu 等人提出了 SSD 算法，引入了多尺度特征图检测技术，使得检测器的精度得到了极大提高。

从上面的介绍我们得出结论：传统方法的目标检测通常需要手动设计特征和分类器，这些方法在处理复杂的场景时表现不佳，并且需要大量的专业知识和时间来调整参数。相比之下，基于深度学习的目标检测技术利用深度神经网络自动学习特征和分类器，具有更高的准确性和鲁棒性，适用于各种复杂场景。目标检测是目标跟踪的前提，我们选择YOLOv5作为研究项目的目标检测器，主要原因有以下几点：

1. 高精度：YOLOv5是目前最高精度的目标检测器之一，能够在多个数据集上实现最先进的性能。
2. 高效率：YOLOv5的设计旨在提供高效的推理速度和小的模型大小，使得它可以在移动设备和嵌入式系统上快速运行。
3. 简单易用：YOLOv5的代码库易于使用，可以轻松地进行训练、测试和部署，即使是对于初学者也非常友好。
4. 可定制性强：YOLOv5的结构灵活，可以通过更改超参数和网络层来满足各种应用需求。

综合以上优点，YOLOv5成为了一个非常流行的目标检测器，特别适合需要高精度和高效率的应用场景。

### YOLOv5目标检测算法

**YOLOv5算法**

**【**韩晓冰**】**

YOLOv5 是由作者 Glenn Jocher 提出，一共有四种大小模型，分别为YOLOv5s、YOLOv5m、YOLOv5l、YOLOv5x。其中 ：YOLOv5s的卷积层数最少、检测速率最快、检测精度最低；另外的 3个模型，卷积层数从小到大依次排列，随着模型复杂度增大检测速率逐渐降低，检测精度逐渐提高。YOLOv5 结构包括四部分：Input、Backbones、Neck、Prediction，如图1所示

图示

描述已自动生成

Input端包含 Mosaic 增强、动态锚框、自适应图片处 理等；Backbones 包括跨阶段局部网络 CSP、空间金字塔池化SPPF，前者有助于减少计算量，后者提高检测精确度；Neck 采用了 FPN+PAN 的结构，下采样增强语义信息，上采样增强定位信息；Prediction 是最后检测输出端。

为了追求更快的速度，我们选用YOLOv5开源网站【注释】发布的预训练好的yolov5s模型。

## 多目标跟踪

多目标跟踪的主要目的是对视频中的多个目标同时进行轨迹跟踪，并维持对应的身份信息。随着近年来多目标跟踪研究的不断增多，其相关方法层出不穷，当前主流的方法可分为以下两类：基于目标检测的多目标跟踪和联合检测和跟踪的多目标跟踪。

许多人在草地上放风筝

描述已自动生成

### 基于目标检测的多目标跟踪 (Tracking By Detection, TBD)

在基于目标检测的多目标跟踪算法中，目标检测和特征提取被分为两步，首先对当前帧的图像执行目标检测，然后提取检测框的特征表示，最后将检测框与已有的运动轨迹进行关联。在 2016 年，Alex 等人[11]提出了一种在线和实时跟踪方法（Simple Online and Realtime Tracking, SORT））利用卡尔曼滤波预测每条跟踪轨迹在当前帧中的运动状态，将预测值与当前帧中候选检测框的交并比作为相似度矩阵得分，最后通过匈牙利算法解决轨迹和候选检测框间的关联任务，这一创新性工作为多目标跟踪研究找到了新的方向。SORT在数据关联时只使用目标的运动信息，因此在遇到频繁遮挡时容易出现身份信息交换的情况。Wojke 等人[12]提出的 DeepSORT 算法解决了该问题，在 DeepSORT 中使用结合运动和外观信息的联合度量来完成数据关联，大大减少了行人身份信息变化的次数，还引入级联匹配解决遮挡导致的概率弥散问题。另外，为解决检测器漏检、误检等问题， Chen 等人[13]设计了一个基于卷积神经网络的评分函数，利用该函数从检测器和跟踪器的输出中生成最优的候选框，进一步拓展了基于检测的多目标跟踪算法。

### 联合检测和跟踪的多目标跟踪（Join Detection and Tracking, JDT）

联合检测和跟踪的多目标跟踪算法采用了多任务联合学习的思想，该类算法通常训练一个可以同时输出目标检测框位置和目标外观特征的模型，提高了多目标跟踪的速度。Wang 等人[14]提出将行人重识别模型合并到 YOLOv3 检测器中， 并设计一个联合检测框回归、anchor 分类以及嵌入学习的损失函数完成同时学习外观特征表示和目标检测的任务。作为第一种联合检测和特征提取的方法，该算法存在很多不足，比如外观信息中缺少多层特征的融合、基于锚框的检测器不适合提取目标外观特征等问题。为了解决这些问题，Zhang 等人[15]提出了 FairMOT，使用深层聚合网络来实现不同尺度的检测，以便提取多层次的行人重识别信息， 同时使用了基于关键点的检测方式，缓解了目标中心偏移问题。此外，研究者们还提出将数据关联也统一到单个网络中，如 Zhou 等人[16]提出的 CenterTrack，通过将检测模型应用于相邻图像帧，结合前一帧的目标检测结果来预测当前帧的目标运动状态并完成数据关联，巧妙地将数据关联问题转为相邻帧之间的检测框回归问题。Peng 等人[17]提出了链式跟踪算法（Chained-tracker），业内首创统一目标检测、特征提取、数据关联三个任务的端到端多目标跟踪网络，并设计了联合注意力模块，相较于其他算法集成度更高。

综上所述，联合检测和跟踪（JDT）和基于目标检测的多目标跟踪（TBD）都是针对视频中出现的多个目标进行跟踪的算法。其中，JDT算法将目标检测和跟踪过程结合在一起，先通过目标检测得到目标的位置信息，再在下一帧图像中通过匹配和预测得到目标的轨迹；而TBD算法则是首先利用目标检测算法检测出所有目标，然后利用某些模型或者算法进行轨迹匹配和更新。相比较而言，JDT算法更加高效准确，但是在目标密度较大时，算法的计算量会变得非常大，从而降低了跟踪的精度。考虑到零售场景会出现人员密集的情况，我们在研究项目中使用DeepSORT作为目标跟踪算法，它能有效解决人员密集的问题。DeepSORT使用卷积神经网络来提取目标的特征，然后利用卡尔曼滤波和匈牙利算法来进行多个目标的跟踪，并且具有较高的准确性和鲁棒性。此外，DeepSORT还具有较好的扩展性和可调节性，可以根据不同的应用场景进行优化和改进。

### DeepSORT目标跟踪算法

**【**郑繁亭**】 【韩晓冰】**

DeepSort是对 Sort追踪算法的改进，为了降低跟踪目标身份编号切换（ID switch）次数，DeepSORT引入了外观特征和级联匹配。DeepSORT的主要目的是生成视频流中多个人物运动的连续轨迹，其总工作流程如下图。DeepSORT算法以目标检测网络检测结果的检测框坐标和置信度作为输入，通过卡尔曼滤波算法对下一帧行人的位置进行预测，再进行级联匹配，然后利用匈牙利算法进行数据关联，最后进行卡尔曼滤波更新。

图示, 示意图

描述已自动生成

DeepSORT计算相似度利用了目标的运动信息和外观信息。对于运动信息，用马氏距离判断预测目标与检测目标的关联度。马氏距离表达式为：

式中，是检测框j的位置；是跟踪器i预测的位置；是检测与预测位置的协方差矩阵。

当目标长久遮挡或视角抖动，则要引入外观信息，通过余弦距离来解决因遮挡带来身份切换的问题。

余弦距离表达式为：

式中，是检测框的本征向量；是跟踪器i对应最近100帧本征向量的集合；为外观特征向量库。

为了充分利用两种信息，采用线性加权的方式求和：

式中：为超参数，当且仅当度量值存在于和之间，才认为目标关联。

由于卡尔曼滤波算法是基于线性匀速运动的，所以DeepSORT算法只能对线性环境下的目标进行状态预测，在行人轨迹非线性的时候也许不能很好地进行预测。

日程表

描述已自动生成

图示

描述已自动生成

## 行人重识别

【来源：<https://zhuanlan.zhihu.com/p/456060221】>

行人重识别（Person Re-identification也称行人再识别，简称为ReID，是近几年智能视频分析领域兴起的一项新技术。它旨在解决这样的问题：给定一个监控行人图像，检索多个视频中的该行人图像。在监控视频中，由于相机分辨率和拍摄角度的缘故，通常无法得到质量非常高的人脸图片。当人脸识别失效的情况下，ReID就成为了一个非常重要的替代品技术。ReID有一个非常重要的特性就是跨摄像头，所以学术论文里评价性能的时候，是要检索出不同摄像头下的相同行人图片。

行人重识别任务主要包含特征提取和相似度度量两个步骤. 传统的方法思路为手工提取图像 特征, 例如: 颜色、HOG (Histogram of oriented gradient)[4]、SIFT (Scale invariant feature transform)[5]、LOMO (Local maximal occurrence) 等。之后, 利用 XQDA (Cross-view quadratic discriminant analysis)[6] 或者 KISSME (Keep it simple and straightforward metric learning)[7] 来学习最佳的相似度度量。然而，传统的手工特征描述能力有限, 很难适应复杂场景下的大数据量任务. 并且, 在数据量较大的情形下, 传统的度量学习方法求解也 会变得非常困难。

近年来, 以卷积神经网络为代表的深度学习在 计算机视觉领域取得了极大的成功, 在多项任务上 都击败传统的方法, 甚至一定程度上超越了人类的 水平[8−9]。在行人重识别问题上, 基于深度学习的 方法可以自动学习出复杂的特征描述, 并且用简单的欧氏距离进行相似度度量便可以取得很好的性能. 换句话说, 深度学习可以端对端地实现行人重识别任务, 这使得任务变得更加简单. 目前, 基于深度学习的行人重识别方法已经在性能上大大超越了传统的方法。这些优势使得深度学习在行人重识别领域变得流行。

【来源：罗浩-基于深度学习的行人重识别研究进展】

述基于深度学习的行人重识别方法根据训练损失可以分为基于表征学习和度量学习，在这里我们介绍这两种不同的学习方法 。

### 基于表征学习(Representation learning)的方法

基于表征学习 (Representation learning) 的方法是一类非常常用的行人重识别方法[3, 25−31] . 虽然行人重识别的最终目标是为了学习出两张图片之间的相似度, 但是表征学习的方法并没有直接在训练网络的时候考虑图片间的相似度, 而把行人重 识别任务当做分类 (Classification) 问题或者验证 (Verification) 问题来看待. 这类方法的特点就是网络的最后一层全连接 (Fully connected, FC) 层输出的并不是最终使用的图像特征向量, 而是经过一个 Softmax 激活函数来计算表征学习损失, 前一层 (倒 数第二层) FC 层通常为特征向量层. 具体言之, 分类问题是指利用行人的 ID 或者属性等作为训练标 签来训练模型, 每次只需要输入一张图片; 验证问题 是指输入一对 (两张) 行人图片, 让网络来学习这两张图片是否属于同一个行人。

分类网络常用的两种损失分别是行人 ID 损失 (Identification loss) 和属性损失 (Attribute loss). 文献 [3, 29] 将每一个行人当做分类问题的一个类别, 用行人的 ID 作为训练数据的标签来训练 CNN 网络, 这个网络损失被称为 ID 损失, 而这种网络被称为 IDE (ID embedding) 网络. IDE 网络是行人重识别领域非常重要的 baseline 基准。

文本, 信件

描述已自动生成

文本, 信件

描述已自动生成

文本, 信件

描述已自动生成

文本, 信件

描述已自动生成

图示, 工程绘图

描述已自动生成

### 基于度量学习 (Metric learning)的方法

度量学习 (Metric learning) 是广泛用于图像检索领域的一种方法. 不同于表征学习, 度量 学习旨在通过网络学习出两张图片的相似度。在行人重识别问题上, 表现为同一行人的不同图片 间的相似度大于不同行人的不同图片。具体为，定义一个映射f(x): ，将图片从原始域映射到特征域，之后再定义一个距离度量函数，来计算两个特征向量之间的距离。最后通过最小化网络的度量损失，来寻找这样的一个最优映射f(x)，使得相同行人两张图片（正样本对）的距离尽可能小，不同行人两种图片（负样本对）的距离尽可能大。而这个映射f(x)，就是我们训练得到的深度卷积网络。

常 用 的 度 量 学 习 损 失 方 法 包 括 对 比 损 失 (Contrastive loss)、 三 元 组 损 失 (Triplet loss)。首先, 假如有两张输入图片 I1 和 I2, 通过网络的前向 传播我们可以得到它们 (归一化后) 的特征向量 fI1 和 fI2 . 之后我们需要定义一个距离度量函数, 这个 函数并不唯一, 只要能够在特征空间描述特征向量 的相似度/差异度的函数均可以作为距离度量函数。通常我们使用特征的欧氏距离或者余弦距离作为度量函数, 即两张图片在特征空间的距离定义为:

手机屏幕截图

低可信度描述已自动生成

三元组损失是最被广泛应用的一种度量学习损失, 之后的大量度量学习方法也是基于三元组损失演变而来。顾名思义, 三元组损失需要三张输入图片. 和 对比损失不同, 一个输入的三元组 (Triplet) 包括一 对正样本对和一对负样本对. 三张图片分别命名为 固定图片 (Anchor) a, 正样本图片 (Positive) p 和 负样本图片 (Negative) n. 图片 a 和图片 p 为一对 正样本对, 图片 a 和图片 n 为一对负样本对. 则三元组损失表示为:



在计算度量损失时, 样本对都是从训练集中随机挑选. 随机挑选样本对的方法可能经常挑选出一些容易识别的样本对组成训练批量 (Batch), 使得 网络泛化能力受限. 为此, 部分学者提出了难样本 采样 (Hard sample mining) 的方法, 来挑选出难样 本对训练网络[37, 41] . 常用的思路是挑选出一个训练 Batch 中特征向量距离比较大 (非常不像) 的正样本 对和特征向量距离比较小 (非常像) 的负样本对来训 练网络. 难样本采样技术可以明显改进度量学习方 法的性能, 加快网络的收敛, 并且可以很方便地在原 有度量学习方法上进行扩展, 是目前广泛采用的一 种技术.

综上所述，度量学习可以近似看作为样本在特征空间进行 聚类, 表征学习可以近似看作为学习样本在特征空 间的分界面. 正样本距离拉近的过程使得类内距离 缩小, 负样本距离推开的过程使得类间距离增大, 最 终收敛时样本在特征空间呈现聚类效应. 度量学习 和表征学习相比, 优势在于网络末尾不需要接一个 分类的全连接层, 因此对于训练集的行人 ID 数量并不敏感, 可以应用于训练超大规模数据集的网络. 总 体而言, 度量学习比表征学习使用的更加广泛, 性能 表现也略微优于表征学习。因此我们选择了基于度量学习的FastReID方法作为研究项目的行人重识别方法，实验证明，\*\*\*\*。

### Fast ReID算法

DeepSORT之所以能有效解决遮挡问题，是因为其搭建了一个重识别神经网络WRN。WRN主要由残差块组成，参数多、计算量大。为了增强重识别网络的速度和识别能力，引入行人重识别模型FastReID替代WRN，可以在实现同样功能的前提下，更\*\*\*，更\*\*\*。

FastReID 的整体由图像预处理、主干网络、特征集成和网络头四大部分组成。整体架构如下图所示：

图示

描述已自动生成

在数据预处理中，FastReID 使用了 Auto-augment 操作，也就是 automl 中的技巧，使用该技巧去实现有效的数据增强，提高特征的鲁棒性。

在 Backbone 中，实现了 ResNet特征提取模型。而在特征集成部分，则实现了最大池化、平均池化、Gem 池化和 Attention 池化四种池化操作。

另外作者实现了四种常见的loss，包括：

* 交叉熵损失
* Arcface loss
* Circle loss，旷视今年提出的
* Triplet loss

在训练时候，作者使用了当今流行的热身训练法（warm up）、余弦衰减（consine decay）等策略。

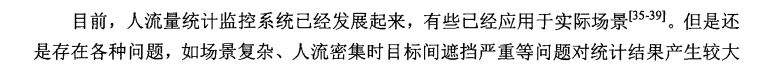
## 人流量统计

【沈爽】

文本

描述已自动生成

国际上许多公司在人数统计应用产品方面进行了大量研宄 与 实验 ， 如 ， 飞 瑞斯 公 司 的 人流量统计 Ｗｉ ｓ ｅ Ｃ ｏ ｕｎｔ 系 列 产 品 ， 统计准 确 度超过 ９ ５ ％ ，可以一 天不 间 断 的 工 作 ， 在 实 际应 用 中 有 很 高 的 实 时 性和 统计准确度。



影响，因此仍需要对监控系统进一步完善。针对商量客流量统计这一课题，我们研究了一种基于卷积神经网络的算法，可以记录每个客人进出商店的具体时间，用于商业决策和分析。

## 本文的主要研究内容

本文的主要研究内容是设计一种YOLOv5 算法、DeepSORT 跟踪算法和fastreid行人重识别算法相结合的方法，以实现商店客流量的统计。我们通过对行人检测、跟踪和重识别算法的优化，实现了对不同角度和光照条件下的顾客进行准确的跟踪与计数，并开发了一个面向商店的客流量统计系统。