博士学位论文

基于自适应特征增强的行人目标检测方法

作者姓名：张天亮

指导教师：叶齐祥　教授

中国科学院大学

学位类别：工学博士

学科专业：信号与信息处理

培养单位：中国科学院大学电子电气与通信工程学院

2020 年 8 月

Pedestrian and Object Detection based on Adaptive Feature Enhancement

A dissertation submitted to

University of Chinese Academy of Sciences

in partial fulfillment of the requirement

for the degree of

Doctor of Philosophy

in Signal and Information Processing

By

Tianliang Zhang

Supervisor: Professor Qixiang Ye

School of Electronic, Electrical and Communication Engineering

August 2020

中国科学院大学
研究生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的指导下独立进行研究工作所取得的成果。尽我所知，除文中已经注明引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的研究成果。对论文所涉及的研究工作做出贡献的其他个人和集体，均已在文中以明确方式标明或致谢。

作者签名：张天亮

日　　期：2020 年 8 月 15 日

中国科学院大学
学位论文授权使用声明

本人完全了解并同意遵守中国科学院有关保存和使用学位论文的规定，即中国科学院有权保留送交学位论文的副本，允许该论文被查阅，可以按照学术研究公开原则和保护知识产权的原则公布该论文的全部或部分内容，可以采用影印、缩印或其他复制手段保存、汇编本学位论文。

涉密及延迟公开的学位论文在解密或延迟期后适用本声明。

作者签名：张天亮　　导师签名：叶齐祥

日　　期：2020 年 8 月 15 日　　日　　期：2020 年 8 月 15 日

摘　要

行人检测是计算机视觉的一个重要研究领域，其旨在从图像或视频中准确定位行人目标，输出其位置和分类置信度。经过近二十年的深入研究和发展，行人检测技术已经在无人驾驶系统、智能监控系统和交通辅助系统等方面得到了初步应用。

尽管行人检测技术已经获得了长足的发展，但仍存在很多问题，包括检测结果的可靠性、模型的复杂性、目标-特征匹配的准确性等。如在交通场景中，由于存在很多前景与背景的相互遮挡，检测结果并不是十分令人满意。为了解决这些遮挡问题，分类器集成方法和分块模型经常被采用，但是这种方法效率低、计算量大、检测精度难以满足需求。其主要原因在于一个固定的特征提取方法很难对遮挡、密集行人目标提取自适应的特征。除此之外，在目标-特征匹配方面，之前的行人检测方法均是通过空间对齐的方式来进行正反例样本的分配，当目标判别性最强的区域偏离目标中心时，分配准则将限制学到特征的表达能力。

因此针对上述问题，本文基于深度网络系统从网络结构、特征校准、特征优化三个视角研究了针对行人与通用目标的自适应特征增强方法。本文的主要研究内容和贡献如下：

(1) 提出一种用于自适应特征增强的深度学习网络结构，称为环状循环网络（CircleNet）。该网络通过扩展特征金字塔网络结构，增加一条由浅层到深层的通路，将浅层到深层和深层到浅层的两条通路组合构成环状结构，通过权值共享可以将它视为循环网络。该网络通过往复式的特征适配，提取更具有表达性的行人特征，该特征在保持高分辨率的同时也具有更强的语义信息。人在观察遮挡和小目标的时候，往往需要进行多次识别。所提出的环状循环网络也在每次循环中对图像中的行人进行多次检测。同时结合行人实例分解训练策略，使得环状循环网络的潜力得到了发挥，在一般行人和遮挡行人上提升了行人目标检测的准确性。

(2) 提出一种解决遮挡行人检测的特征增强方法，称之为特征校准（Feature Calibration, FC）模块。首先我们提出行人激活模式概念，该模式是深度卷积特征每个通道学到的行人局部特征。通过聚合所有激活模式可以得到行人激活图，然后使用该激活图加强图像级别的特征，进行像素级特征校准。像素级特征校准

可以加强行人可见部分特征，并且抑制来自遮挡区域的噪声干扰。这是解决遮挡行人检测问题的关键。在提取目标级的区域特征之后，使用目标级特征校准进行特征增强。目标级特征校准融合了自适应的上下文信息，可以学习背景和行人的共生信息。本文的方法可以适应性地根据遮挡情况加强或减弱特征，最终在不同遮挡情况下提升行人目标检测的鲁棒性。

(3) 提出了特征选择-抑制的特征增强算法。传统的视觉目标检测算法一般通过直接的目标-特征匹配进行模型学习。然而，针对有些倾斜或遮挡的目标，传统算法并不能差异地看待不同锚点框在训练中的贡献程度。我们提出的算法引入了锚点包学习策略，并将目标-特征匹配准则由手工指定改为动态自适应，实现了目标-特征的优化匹配。考虑到传统的多示例学习方法容易产生模型的局部最优解，本文提出了“选择-抑制-增强”的对抗训练策略。该策略通过扰动得分最高的锚点框的特征来多次降低其置信度，从而让网络考虑更多的锚点框。其本质是增加更多的低置信度但位置正确的锚点框参与学习的机会，从而缓解陷入局部最优的情况，实现特征的自适应增强。

本文的研究为视觉行人目标的表征提供了新思路，相关研究方法对于通用视觉目标检测具有借鉴意义。本文研究的自适应特征增强开拓了深度学习领域的新方向。

关键词：行人目标检测，视觉目标检测，特征增强环状循环网络，行人激活图，特征校准，多示例学习

Abstract

Pedestrian detection is a key problem in computer vision, which aims to identify pedestrians from images or videos and outputs location and classification confidence. After nearly two decades of research and development, pedestrian detection technology has been initially used in unmanned driving systems, intelligent monitoring systems, and traffic assistance systems.

Although the research on pedestrian detection has obtained great development, it is still not enough reliability and stability. Especially in traffic scenes, there are too many occlusions. In order to solve these occlusion problems, classifier integration methods and part models are often used, but these methods are inefficient and require a large number of computing resources. Detection accuracy is also difficult to meet the requirements. The main reason for this is that a fixed feature extraction method is difficult to adapt to occluded and dense pedestrians. In addition, the previous pedestrian detection methods all assign positive and negative examples through spatial alignment. When the most discriminative region deviates from the object center, the assignment criterion will limit the representative ability of learned features.

To solve above problems, this paper proposes several adaptive feature enhancement methods for pedestrian and general object detection from three perspectives: network structure, feature calibration and feature optimization. The main research contents and contributions of this paper are as follows:

(1) Propose a novel network structure for adaptive feature enhancement, referred to as CircleNet, to achieve feature adaptation by mimicking the process humans looking at low resolution and occluded objects: focusing on it again, at a finer scale, if the object can not be identified clearly for the first time. CircleNet is implemented as a set of feature pyramids and uses weight sharing path augmentation for better feature fusion. It targets at reciprocating feature adaptation and iterative object detection using multiple top-down and bottom-up pathways. To take full advantage of the feature adaptation capability in CircleNet, we design an instance decomposition training strategy to focus on detecting pedestrian instances of various resolutions and different occlusion levels in each cycle. Experiments show that CircleNet improves the performance of occluded and low-resolution pedestrians.

(2) Design a feature enhancement method to solve the occlusion pedestrian

detection, which is called the Feature Calibration (FC) module. First, the concept of pedestrian activation pattern is proposed, which is the local features of pedestrians learned by each channel of the deep convolution feature. Pedestrian activation maps can be obtained by aggregating all activation patterns, and then the activation maps are used to enhance image-level features and perform pixel-wise feature calibration. Pixel-wise feature calibration highlights the visible parts and suppresses the occluded parts of pedestrians. This is the key to solving the problem of occlusion pedestrian detection. After extracting object-level features, we use region calibration to enchance object features. It fuses adaptive context information and learns the concurrence information of background and pedestrians. Our method can adaptively strengthen or weaken the features according to the occlusion conditions, and detect occluded pedestrians robustly.

(3) Propose a feature enhancement algorithm based on multiple instance learning, referred to as multiple anchor learning (MAL). Traditional object detection algorithms generally use direct object-feature matching for model learning. However, for some tilted or occluded objects, the contribution of different anchors in training can not be viewed differently. The proposed algorithm introduces an anchor bag, and changes the anchor selection from static to dynamic. It achieves the object-feature optimal matching. Our approach selects the most representative anchors from each bag. Such an iterative selection process is potentially NP-hard to optimize. To address this issue, we solve MAL by repetitively depressing the confidence of selected anchors by perturbing their corresponding features. In an adversarial selection-depression manner, MAL not only pursues optimal solutions but also fully leverages multiple anchors/features to learn a detection model.

The research in this paper provides some new ideas for the representation of visual pedestrian, and these methods are useful guides for general object detection. The adaptive feature enhancement research in this paper opens up a promising direction for deep learning.

Key Words：Pedestrian Detection, Object Detection, CircleNet, Pedestrian Activation Map, Feature Calibration, Multiple Instance Learning

目录

图目录

表目录

第1章　绪论

1.1 引言

自进入 21 世纪以来，我国信息技术产业在生产和科研方面快速发展。人工智能最近成为了信息技术产业中一个热点研究方向[1][2][3][4],它可以服务于人类的各行各业，并且具有创造巨大经济与社会价值的潜力。人工智能虽已被提出许久，但是很少被应用到现实生活技术中。但随着计算机计算资源和计算机硬件条件的提升，算法的性能得到了极大的改善，它也再次回归到大家视野中。

计算机视觉是人工智能领域的一个重要分支，就像视觉系统是人类信息感知的主要来源，对于计算机来说，计算机视觉系统是计算机智能感知的重要手段，其目标是把类似人类感知和理解图像语义信息的能力赋予计算机[5][1][6][7]。它不仅代表着科学技术发展的前沿，同时也是带领人们走向人工智能时代的必经之路。面对当今的大数据时代，快速有效地在海量数据中实时对目标进行解析是非常重要的。行人检测技术就有这样的能力，它也是计算机视觉关键性技术之一，它可以应用于交通系统中人流和车流监控、居民区和公司的安保监控、医院的医护监控场景等。在车辆和行人的目标发现、检测和识别，人流和车流量估计，人群密集地区拥挤程度估计与控制和一些重要物品的防盗中，计算机视觉技术发挥着重要作用。

1.2 课题研究背景与意义

1.2.1 课题的背景

人工智能（Artificial Intelligence，AI）的迅速发展正在改变人们的生活方式。为抓住人工智能发展的重要战略机遇，构建我国人工智能发展的优势，加快建设创新型国家和早日成为世界科技强国，2017 年我国制定了新一代人工智能发展规划[8]。人工智能技术已经成为国际各个大国之间竞争的焦点，它将成为引领未来的战略性关键技术。世界主要发达国家把发展人工智能作为提升国家综合国力的重大战略。最近十年，随着计算资源的蓬勃发展，人工智能也迎来了第三次热潮。越来越多人工智能的技术被应用于现实生活中，极大地改善了人们生活质量。

计算机视觉[9]（Computer Vision，CV）是一个跨学科的科学领域，它可以解

决计算机如何从数字图片和视频中获取高层理解的问题。从工程的角度来看，它模拟完成人类视觉系统可以完成的任务。这是一个不断发展成熟起来的新兴领域，在今天该领域已经在理论和实际应用中取得了丰硕的成果。随着深度学习[10]（Deep Learning，DL）方向的成熟，图像和视频的信息获取速度正在以前所未有的速度不断发展。至今，计算机视觉在医学成像[11]、地球资源遥感检测[12]、天文学[13]、工业自动化等诸多领域的应用已经越来越多，它正在不断推进着人类的进步和发展。而在计算机视觉研究领域下，有一个受到学术界和工业界关注的热门方向，即行人检测。

行人检测是计算机视觉的一个关键问题[14][15][16][17]，其中一些行人检测的应用对生活质量产生着积极的影响。行人检测就是计算机对于输入的图像和视频判断其中是否包含行人，如果包含行人，需要给出行人的具体坐标位置和置信度。人是人机交互环境中最主要的组成部分，因此检测和跟踪人也是现代工程学中最具有潜力和最具挑战的任务之一。行人检测也是行人跟踪[18]、行人再识别[19]、行为分析[20]、身份识别[21]等问题的基础，同时它也在人们日常生活和城市安防中起到了巨大的作用。

1.2.2 课题的应用领域

行人检测已经在智能视频监控系统、车辆辅助系统和智能机器人系统等多个领域广泛应用。它是自动驾驶[22]、智能视频监控系统[23][24]、车辆辅助系统[25]和智能机器人系统[26]的核心技术。

1、智能视频监控系统

随着智慧城市概念的提出，越来越多的摄像头安装在了公共场所，包括街道、商场、银行、学校和交通枢纽区域等。当前这些系统主要使用存储功能，一旦需要调取监控内容时，主要是由人工搜索完成，而超大的数据量使得搜索监控内容异常繁重。智能的监控系统和其核心的自动化行人检测算法，可以缓解工作人员的工作量，防止由于人员疲惫而漏掉的重要信息，能进一步提升监控系统的准确度和搜索速度。

2、汽车辅助驾驶

根据国家统计局提供的数据[27]，如图 1.1 所示，2015 至 2017 年交通事故死亡人数逐年增加，其中 2017 年交通事故死亡人数总计达 63772 人。而汽车先进

驾驶辅助系统（Advanced Driver Assistance Systems, ADAS）和行人保护系统[25]（Pedestrian Protection Systems, PPSs）的应用可以在一定程度上减少交通事故的发生，避免人员伤亡和财产损失。ADAS 和 PPSs 可以有效地检测和跟踪行人，并为车辆提供必要的分析结果，防止潜在的交通事故，并且可以在不可避免发生事故的情况下降低其严重性。随着传感技术的进步、人工智能和计算机视觉的成熟以及机器计算能力的提高，这些系统的应用会进一步扩大推广规模。

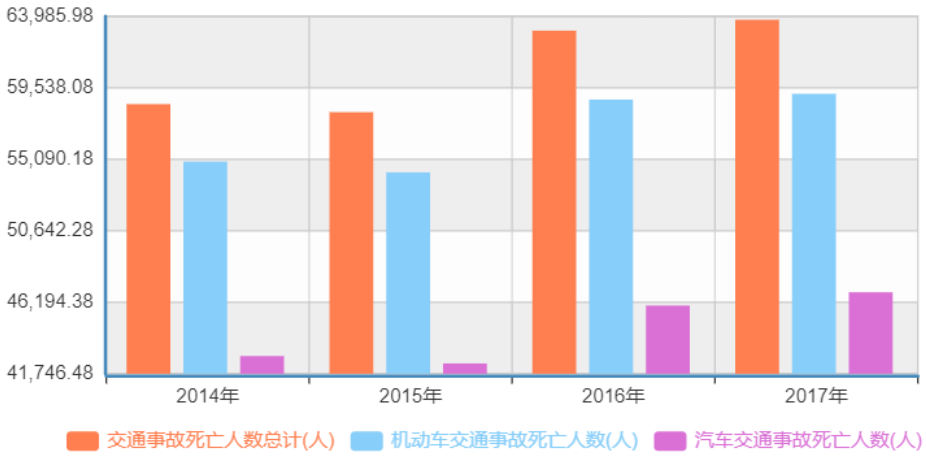


图 1.1 交通事故死亡人数统计[27]

Figure 1.1 Statistics of fatalities in traffic accidents

3、智能机器人

随着硬件和软件性能的提升，咨询机器人、导航机器人、安防机器人等越来越普遍地出现在人们日常生活中。它们可以给大众提供一些基础服务，使得生活更加便利。这些智能机器人是以服务人为核心的，因此识别人对于它们的指令是核心算法。通过行人检测对周围人识别，再通过行为分析或者语音交互完成命令的下达，以完成最终的服务目的。

行人检测还有更多的应用场景，它对改变人们生活生产方式上有着深远的意义和影响。

1.3 国内外研究现状

目前，许多重要的国际学术会议和权威期刊都针对行人检测领域的最新理论研究和应用进展做出了专门讨论，如：International Conference on CVPR（Computer Vision and Pattern Recognition）、ICCV （International Conference on

Computer Vision）、ECCV （European Conference on Computer Vision）、 ACCV（Asian Conference on Computer Vision）、IJCV （International Journal of Computer Vision）、PAMI（IEEE Transaction on Pattern Analysis and Machine Intelligence）、TITS （IEEE Transactions on Intelligent Transportation Systems）、TIP （IEEE Transaction on Image Processing）等将行人检测研究作为主题内容之一，为该领域的研究人员提供了更多的交流机会。

国内的重要研究机构和公司有中科院计算所、中科院自动化所、华为诺亚方舟实验室、滴滴出行科技、百度深度学习研究院（Institute fo Deep Learning，IDL）、地平线机器人、海康威视、商汤科技、公安部第三研究所和图森未来等。其中 IDL 和图森未来在自动驾驶方面已经有相当的积累，百度的自动驾驶出租车队已于 2019 年在长沙市开始试运营，公安部三所和海康威视在智能安防领域中具有绝对的市场占有率[28]。

国外的研究机构主要包括：谷歌（Google）、脸书（Facebook）、苹果（Apple）和微软等。谷歌 X 实验室、苹果和特斯拉都在研发自动驾驶汽车，而行人检测是自动驾驶的关键技术之一。通用公司、沃尔沃公司、特斯拉公司和日产公司等汽车巨头的 L1-L2 级自动驾驶汽车已经开始大规模量产，并且将行人检测系统（Pedestrian Detection System，PDS）加入到了自动驾驶系统中[29]。

通常根据行人检测特定问题把行人检测方法分为通用性行人检测和遮挡行人检测两个部分，下文我们将针对这两个方面分别进行介绍。

1.3.1 通用性行人检测方法

行人检测技术已经发展了数十年。早在 2003 年，Viola 和 Jones 在行人检测任务中就提出了使用 VJ 检测器[30]用于行人检测，由此 VJ 检测器成为了现代行人检测技术兴起的标志。VJ 检测器使用了移动和外表信息，同时使用 AdaBoost 选择特征的一个子集训练分类器。2005 年，Dalal 和 Triggs 引入了具有里程碑意义的 HOG（Histograms of Oriented Gradient）检测器[31]，该检测器后来成为可变行组件模型（DPM）[32]的重要组成部分。2014 年 P. Dollar 等人[33]提出积分通道特征（Integral Channel Feature, ICF），并将其应用于行人检测中，实验表明由增强决策树选择的方向梯度和颜色特征对行人检测很重要，性能明显优于之前的行人检测器。2015 年之前，行人检测方法基本都在使用传统手工设计的特征

[30][31][32][33]，并且结合决策树方法。传统的行人特征提取方法往往根据研究者对问题的理解和先验知识进行设计，这样限制了特征的学习能力。

与此同时，深度学习的研究也正在高速发展。2012 年，加拿大多伦多大学的Hinton教授及其学生Alex使用卷积神经网络（Convolutional Neural Networks，CNN）在 ImageNet 图像分类数据集上的评测性能远远领先于第二名，它将分类任务中前 5 候选错误率由 26%降低至 16%，由此掀起了深度学习的热潮[34]。卷积神经网络消除了手工设计特征提取算法中的主观性和片面性，利用多组卷积核自动从图像中进行特征学习。本文将行人检测方法分为 4 类，分别是传统手工特征、深度学习特征、上下文和尺度信息融合和无锚点框（Anchor-Free）方法。下面我们对每种方法分别进行介绍。

传统手工特征行人检测方法：传统手工特征的设计一般用到模板、关键点、梯度和边缘等信息。早期大多数行人检测方法主要关注特征的提取，包括模板（Templates）、Harr[35]特征、边缘方向直方图特征[36]（Edge Orientation Histograms Features，EOH）、梯度方向直方图特征[31]（Histogram of Oriented Gradients Features，HOG）、Shapelet 特征、局部二值化特征[37]（Local Binary Pattern Features，LBP）、主导方向模板特征[38]（Dominant Orientation Template Features，DOT）、共现特征（Co-Occurrence Features）[39]和协方差特征[40]（Covariance Features）等。

深度学习特征行人检测方法：2012 年，Krizhevsky 等人[34]提出深度卷积神经网络（Deep Convolutional Neural Network, DCNN），并且根据作者的名字进行命名，称之为 AlexNet。DCNN 在大规模视觉识别挑战任务（Large Scale Visual Recognition Challenge, ILSRVC）的图片分类任务上取得了突破性进展。随后掀起了深度学习的高潮，在通用目标检测、边缘检测、图像搜索等多个任务中，众多深度学习模型被提出使用。行人检测方向也随即出现了少量使用深度特征的方法。2015 年 Jan 等人[41]提出使用深度卷积网络来提取行人特征，达到了和传统手工特征相同优秀的性能。同时作者分析了网络结构的选择、参数和训练数据的影响。Tian[42]提出应用 ACF 检测器[33]产生候选区域，然后使用一个深度网络学习行人属性和场景属性，最后联合优化语义分割和行人检测任务。之后，很多研究者开始采用手工设计特征和深度特征结合的方式。典型的有 RPN+BF[44]（Region Proposal Network and Boosted Forests），它研究了 Fast/Faster R-CNN[43][45]检测框架在行人检测中存在的问题，提出采用增加深度特征的分辨率处理小尺度

行人，同时结合级联的 Boost Forest 挖掘难反例样本。Zhang[46]等经过一系列手段将原始的 Faster RCNN 改进为 Adaptive Faster RCNN，并应用到行人检测任务中。其中包括：针对行人样本使用了更加精细的卷积特征；重新量化了 RPN 中锚点框的尺度；关注行人数据集标注中忽视区域的处理；使用上采样的图片作为输入；最后将 SGD 求解器替换成 Adam[47]求解器。它在广泛使用的 Caltech[14]行人检测数据集上获得了较高的性能。

使用上下文和尺度信息融合的行人检测方法：PCN[48]（Part and Context Network）是一个部件和上下文网络。它设计了两个分支网络，分别利用了身体部件语义和上下文信息来帮助检测行人。在身体部件分支，身体部件的语义信息可以通过循环神经网络彼此进行沟通；在上下文分支中，采用了局部竞争机制来适应上下文尺度的选择。深层特征和浅层特征融合的方法[44][49][50]被广泛应用于加强特征的判别，从而提升了小目标的检测性能。

无锚点框的行人检测方法：传统行人检测方法是基于滑动窗口的方法来进行提取行人候选区域的，然而最近几年随着深度神经网络的能力逐渐增强，使用神经网络提取目标候选区域的方法被提出，比如经典的 RPN[45]（Region Proposals Network）。RPN 可以生成高质量的候选框，淘汰了传统的滑动窗口方法。但是这些基于深度特征的候选区域生成方法依赖于锚点框（Anchor）的初始化，不同数据集初始化的锚点框的参数需要重新设定，所以一些摒弃锚点框的方法被提出。TLL[49]（somatic topological line localization）是一种躯体拓扑线定位方法，首次将行人检测问题转化为类分割问题，使用全卷积网络生成行人躯体掩膜，最后利用马尔科夫随机场（Markov Random Field, MRF）作为后处理步骤来消除遮挡情况的不确定性。基于 Faster R-CNN 的方法需要复杂的锚点框参数配置，但 CSP[50]（Center and Scale Prediction）可利用一个简单的框架直接进行高层语义特征提取，该方法包含特征提取网络和检测网络，其中检测网络实现目标中心点位置检测和目标尺度预测。CSP 简化了基于 Faster R-CNN 框架的行人检测方法，去掉了用于生成行人候选区域的锚点框配置。这些无锚点框的新颖的方法值得学者们继续研究。

1.3.2 遮挡行人检测方法

在交通和商场等场景中，有很多行人结伴而行的情况，所以存在很多的遮挡

问题。为了解决这些遮挡问题，一种简单的做法就是使用分类器集成方法，即人工定义几种遮挡模式，针对每一种模式使用一个单独的分类器进行处理。这种方法效率低，计算量大。另外一种做法根据先验知识把行人分成不同的部分，每个部分的特征单独处理，融合多组块的特征，最后利用融合后的特征进行判别，这种方法被称之为分块模型。分块模型的缺点是不能覆盖每一种遮挡情况，算法复杂度随着块数量的增加而增长。代替分块模型，注意力机制被提出和使用。与分块模型不同的是，注意力机制可以实现像素级的特征加权，不需要采用多组模块，因此在计算量上会优于分块模型。以下部分将详细介绍这些方法。

分类器集成的行人检测方法：Franken-classifiers[51]训练了许多特定遮挡模式的分类器，在遮挡行人检测上取得了性能的提升。PDOE+RPN[52]通过同时回归行人全身位置和行人可见部分位置来完成行人检测和对遮挡程度的估计，除此之外还更新了正例样本的选取准则，即目标候选区域与标注样本的重叠面积和标注目标可见区域的重叠面积同时满足大于设定的阈值。

分块模型的行人检测方法：OR-CNN[53]（Occlusion-aware R-CNN）是一种遮挡感知的 R-CNN，可用来解决遮挡行人检测问题。该方法设计了一个聚合损失函数（Aggregation Loss，AggLoss），强迫目标候选区域相互靠近，并且紧致的定位到相应的目标。除此之外，根据人体先验知识，它把人体分为 5 个部分，每个部分经过一个遮挡处理单元。如果该区域被遮挡，遮挡处理单元屏蔽该块的特征，最终使用所有块的融合特征作为目标特征。

注意力机制的行人检测方法：SDS-RCNN[54]（Simultaneous Detection and Segmentation R-CNN）将分割注入网络，联合监督语义分割和行人检测任务。通过附加的语义分割监督信息，它引导了共享层的特征表达。SSA-CNN[55]（Semantic Self-Attention CNN）在 SDS-RCNN 的基础上，将语义分割监督扩展到多层，最后将多尺度层的语义注意力信息结合到卷积神经网络中以增强行人检测。Faster-RCNN+ATT[56]（FasterRCNN+Attention）利用三种注意力机制来解决行人检测遮挡问题，其中包括自注意力网络、可见框注意力网络和部件注意力网络，后两种网络使用了额外的可见区域框信息和人体关键点信息。GDFL[57]（Graininess-aware Deep Feature Learning）是一种颗粒度感知深度特征学习方法，该方法将细粒度的信息融入卷积特征中，使得网络对与行人身体部件更具判别性，还提出一种行人注意力机制，用于识别行人区域，减少背景干扰。

其他方法：相比于通用目标检测，行人检测问题需要面临更密集目标的场景。非极大值抑制（Non-Maximum Suppression, NMS）是检测算法通用的后处理步骤，其目的在于去掉冗余和重叠率高的检测结果。Adaptive-NMS[58]被提出用于解决非极大值抑制阈值敏感的问题，该模型包含一个子网络学习密度得分，再根据行人密度，动态地调整 NMS 的阈值。Repulsion Loss[59]利用了一个新的检测框回归损失函数，这个损失函数驱使检测结果和目标位置相吸引，并且和周围其他干扰目标相排斥。

这些方法都极大地推动了行人检测和遮挡行人检测方向的前进，然而这些方法都采用了一个固定的特征提取方法，难以对遮挡、密集行人目标进行自适应。同时，它们缺少对行人检测主干网络的研究和显式地建模遮挡模式的方法。

1.3.3 存在的问题

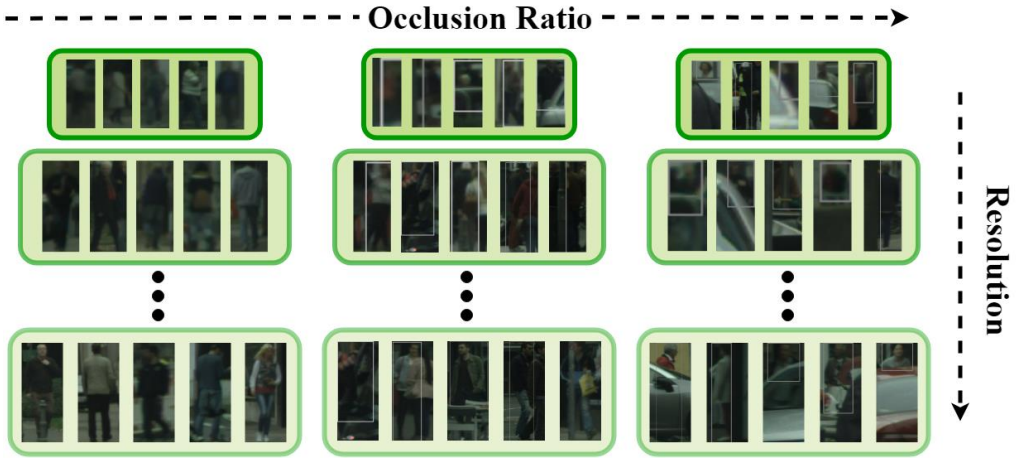


图 1.2 CityPersons 数据集中的行人样本

Figure 1.2 Pedestrian samples in the CityPersons dataset

本论文主要研究内容是计算机视觉中的行人检测问题，即在给定输入视频或图片中，输出行人的具体位置坐标和置信度。在现实场景中，行人检测面临很多的困难，比如复杂的背景、光照变化、人体姿态多样、不同的观察视角等。除此之外，在行人检测的应用场景中，例如视频监控和自动驾驶场景，行人经常聚集在一起并且相互遮挡，或者被车辆或其他障碍物所遮挡。如图 1.2 所示，我们展示了来自 CityPersons[46]数据集的不同遮挡情况和不同分辨率的一些行人样本。在公开数据集 CityPersons 的验证集合中，总共有 3157 个行人样本，其中 48.8% 的行人存在相互遮挡问题[59]，这些问题给行人检测带来了巨大的挑战。如何检测

这些遮挡和低分辨率的行人是当前行人检测中的一个难题。

1.4 本文的研究内容与主要贡献

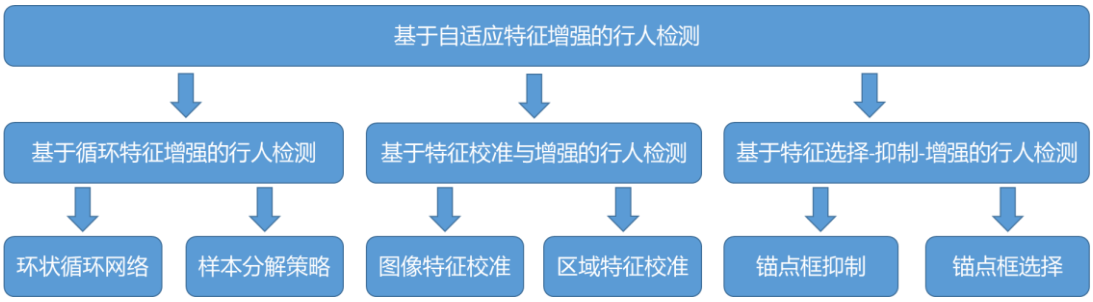


图 1.3 本文研究内容

Figure 1.3 Research content of this article

本文的研究内容如图 1.3 所示，对于遮挡行人检测问题，从主干网络设计、行人特征增强和锚点框匹配及优化三个方面，分别解决了遮挡行人检测中存在的行人多尺度、遮挡行人多模式、遮挡样本的锚点框匹配存在差异等问题。我们提出环状循环网络提取主干特征、特征校准模块（像素级图像特征校准和目标级区域特征校准）和基于多示例学习的锚点框学习训练策略，旨在提升图像特征和目标特征的表示能力，同时使算法适应遮挡行人检测。研究内容包含以下几个方面：

(1) 提出一种深度学习主干网络，称为环状循环网络（CircleNet）。该网络通过扩展特征金字塔网络结构，增加一条由浅层到深层的通路，将浅层到深层的通路和深层到浅层的两条通路组合构成环状结构，通过权值共享可以将它视为环状循环网络。该网络通过往复式的特征适配，提取更有表达性的行人特征，该特征在保持高分辨率的同时具有更强的语义信息。人在观察遮挡和小目标的时候，往往需要多次进行识别。我们提出的环状循环网络也在每次循环中对图像中的行人进行多次检测。同时结合行人实例分解训练策略，使环状循环网络的潜力得到了进一步发挥，在一般行人和遮挡行人上都提升了检测的准确性。

(2) 提出一种解决遮挡行人检测的特征加强方法，称之为特征校准（Feature Calibration, FC）模块。首先我们提出行人激活模式概念，该模式是深度卷积特征每个通道学到的行人局部特征，比如脚步、手臂和头部等。再聚合所有激活模式，即可得到行人激活图。然后使用该激活图加强图像级别的特征，进行像素级的特征校准。像素级的特征校准可以加强行人可见部分的特征，并且抑制来自遮

挡区域的噪声干扰。这是解决遮挡行人检测问题的关键。在提取目标级特征之后，我们使用目标级特征校准，该模块融合自适应的上下文信息，可以学习背景和行人的共生信息。我们的方法可以适应性的根据遮挡情况加强或减弱特征，在不同遮挡率的情况下都可以鲁棒地检测行人。

(3) 提出特征选择-抑制的特征增强算法，称为多锚点框学习（Multiple Anchor Learning, MAL）。该算法基于多示例学习方法引入锚点包学习策略，将锚点框的训练由静态改为动态，并且随着训练的迭代动态调整锚点包的尺寸。传统的锚点框学习算法中，锚点框正反例的分配只考虑锚点框和标注框之间的交并比，当大于一定阈值时即视为正例训练样本，并且在训练过程中正负样本的分配从始至终都不会变化。然而更合理的设计应该是不同锚点框在训练中的贡献程度应该是不同的，比如对于一些倾斜或遮挡的样本。我们提出的锚点包模型选取了前k个锚点框作为正例包，动态调整包的尺寸，并且参照多示例学习的方法进行求解。除此之外，我们提出使用锚点框“选择-抑制-增强”的对抗训练策略，它可以缓解在优化锚点框时陷入到局部最优的情况。

1.5 本文的组织结构

第一章，绪论。主要论述行人检测算法的研究背景和研究意义，总结了国内外行人检测的研究现状，归纳了当前行人检测的主要分类及方法，介绍了几类经典的遮挡行人检测解决方案，分析了当前研究存在的难点问题，明确了本文的主要研究目的和研究内容，列出了本文的主要贡献。

第二章，相关工作与技术。首先介绍了通用目标检测方法，包括双阶段目标检测器和单阶段目标检测器。详细介绍了与本文相关的双阶段目标检测器的主干网络、候选区域提取网络和感兴趣区域池化等概念。然后介绍了一般性行人检测方法和遮挡行人检测相关方法。最后介绍了几个应用广泛的行人检测数据集和行人检测算法评估指标。

第三章，基于循环特征增强的行人检测。首先，我们介绍了相关的特征金字塔网络和路径汇聚网络，提出环状循环网络。然后，针对环状循环网络的结构和往复式的特征适应能力进行分析。之后提出行人分解训练策略和行人注意力区域辅助训练方法。最后进行实验验证、参数选择和检测结果可视化。

第四章，基于特征校准与增强的行人检测。我们提出了特征校准网络结构与

行人激活模式的概念，介绍了行人激活图的生成方式。结合行人激活图，我们提出了基于图像级别的像素特征校准和基于目标级别的区域特征校准模块。最后对该模型进行了实验分析和验证。

第五章，基于特征选择-抑制-增强的视觉目标检测。首先，我们提出了多锚点框学习概念，然后回顾了 RetinaNet 中锚点框的学习方式。其次介绍了多锚点框学习损失函数和“选择-抑制-增强”对抗优化训练策略。最后在通用目标检测数据集上进行了实验验证，并且分析了各个部分对检测结果造成的影响。

第六章，总结与展望。总结了本文的主要内容，并对未来的工作进行了展望，还探讨了如何进一步提高行人检测的鲁棒性和准确性等研究热点和难点。

第2章　相关工作与技术

在近十几年的时间里，行人检测算法得到了长足的发展。特别是在深度学习算法出现之后，一些行人检测算法在部分工业方向得到了初步应用。在图 2.1 中，我们列出了 2012 年以来主流的行人检测方法，其中红色的名称表示几个重要的行人检测数据集。在本章中，我们首先介绍了通用目标检测方法，包括双阶段目标检测器和单阶段目标检测器。详细介绍与本文相关的双阶段目标检测器的主干网络（Backbone）、候选区域提取网络和感兴趣区域池化等概念。然后介绍了一般性行人检测研究方法和遮挡行人检测相关方法。最后介绍了几个应用广泛的行人检测数据集和行人检测算法评估指标。

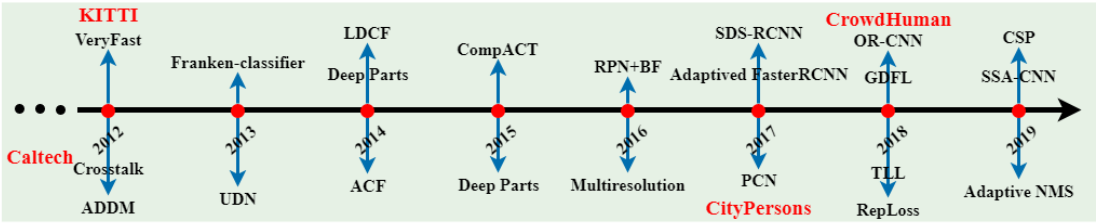


图 2.1 近期主要行人检测方法总览

Figure 2.1 Overview of recent major pedestrian detection methods

2.1 通用目标检测方法

通用目标检测器可以主要分为双阶段目标检测器和单阶段目标检测器两类。双阶段目标检测器性能高，但单阶段目标检测器网络结构简单、速度快。接下来我们对这两个方向的经典工作进行介绍。

2.1.1 双阶段目标检测方法

Faster R-CNN[45]是双阶段目标检测器的经典工作。它是在 Fast R-CNN[43]（Fast Region-based Convolutional Neural Networks）基础上发展而来的，通过增加候选区域提取网络（Region Proposal Network, RPN）替代了原来的目标候选区域提取方法，比如 Selective Search[60]、EdgeBoxes[61]和 MCG[62]等。Selective Search 算法耗时较长，在 CPU 上以 0.5 帧的速度处理图片，并且它不能融入到 Fast R-CNN 网络中，因此整个框架无法端到端的运行。Faster R-CNN 的网络结构如

图 2.2 所示。由主干网络、候选区域提取网络、感兴趣区域池化层、全连接层、分类子网和回归子网构成。我们将对这里的关键技术分别进行介绍。

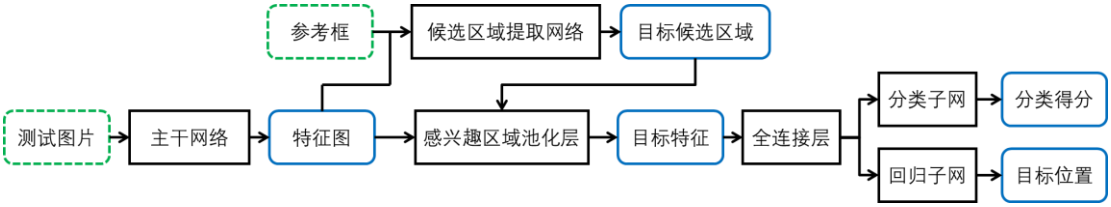


图 2.2 Faster R-CNN 流程图

Figure 2.2 The framework of Faster R-CNN

2.1.1.1 主干网络

VGGNet[34]：Faster R-CNN 的第一个模块是使用一个深层全卷积网络作为主干网络，即 VGG16。VGGNet 曾经在 ILSVRC2014 上获得定位比赛第一名和分类比赛第二名。它还有另外一个结构 VGG19，由于 VGG16 结构更加简单，参数量相对少，所以被广泛使用。去掉原始的 VGG16 的全连接层和最后一个池化层，只使用 Conv1-Conv5 作为主干网络，可用于图像的特征提取。VGG16 由 13 个卷积层、5 个池化层组成。VGG16 的网络结构示意图如图 2.3 所示，图中将具有相同特征分辨率的卷积层定义为一组，一共有 5 组。因为浅层的特征学习到的是纹理信息，在预训练中浅层卷积核已经得到了很好的训练，所以在 Faster R-CNN 训练的过程中，需要冻结浅层 Conv1-Conv2 的参数，不进行微调，只训练后半部分。

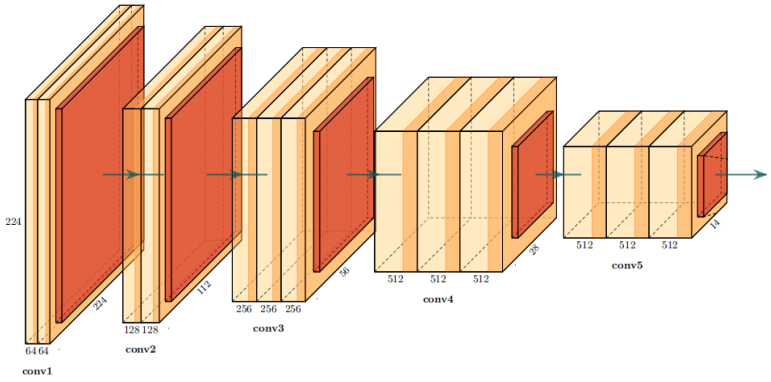


图 2.3 VGG16 网络结构示意图

Figure 2.3 Illustration of VGG16 architecture

GoogLeNet[63]：它是 2014 年 Christian Szegedy 提出的一种全新的深度学习结构，并且在 ILSVRC2014 分类比赛中获得第一名。GoogLeNet 使用的核心模块是 Inception 模块，如图 2.4 所示。整个网络就是由多个 Inception 模块串联构成。Inception 模块的主要贡献是在同一层使用不同尺度的滤波器以保留更多更丰富的空间信息，同时使用1×1的卷积进行升降维，这样较少了网络参数量，不易发生过拟合。

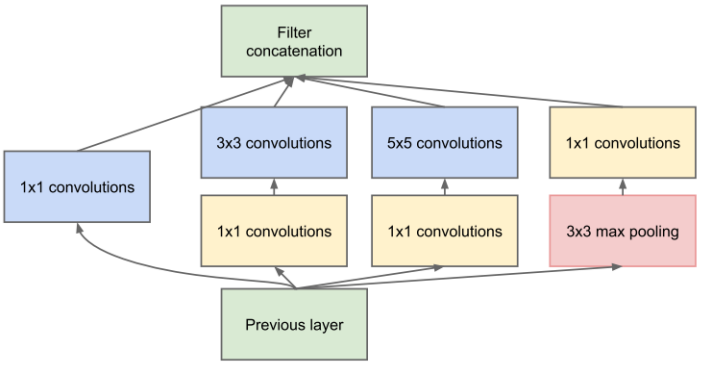


图 2.4 Inception 模块[63]

Figure 2.4 Inception module

ResNet[64]：该论文迄今为止已经拥有超过 5 万次的引用量，是最成功的 CNN 模型之一，同时还获得了计算机视觉和模式识别会议（Computer Vision and Pattern Recognition，CVPR）2016 年的最佳论文。在 ResNet 论文中，作者认为网络的每一层不应该学习整个特征空间的变换，只需要对前一层的残差进行修正即可，这样做可以有效地训练更加深层的网络[65]。通过 ResNet 提取的特征具有较强的泛化能力，在 ILSVRC2015 检测和定位比赛中，该网络获得了冠军。考虑到网络的轻便性，我们在后边的研究方法中也是以 ResNet50 作为模型的主干网络。

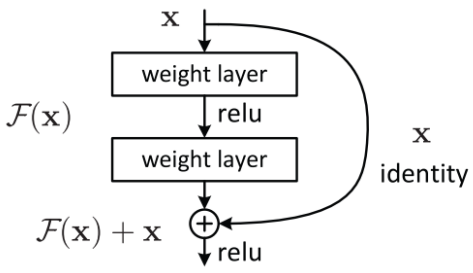


图 2.5 残差学习[64]

Figure 2.5 Residual learning

2.1.1.2 候选区域提取网络

候选区域提取网络（Region Proposal Network, RPN）可以接受任意尺寸的图像作为输入。它输出一组矩形目标候选区域，并且评测每一个区域是目标的可能性。它可以看作是一个简单的单阶段检测模型，并且它的输出是二分类，即只区分目标是前景还是背景。为了生成目标候选区域，在最后共享的卷积层输出的卷积特征图上滑动一个小型的子网络。这个子网络将卷积特征图的 $n \times n$ 空间窗口作为输入。每个滑动窗口都映射到一个较低维的特征（VGG 为 512 维），再后接 ReLU 激活函数。这个特征被送入两个子网络，分别是分类子网络和回归子网络。在论文中使用 $n = 3$，这个特征在原始图像上的感受野比较大，以 VGG 为例，感受野有 228 个像素。由于小子网络以滑动窗口的方式运行，因此全连接层将在所有空间位置上共享。这个结构可以很天然的用 $n \times n$ 的卷积层和两个同级 1×1 卷积层（分别用于回归和分类）实现，网络结构如图 2.6 所示。

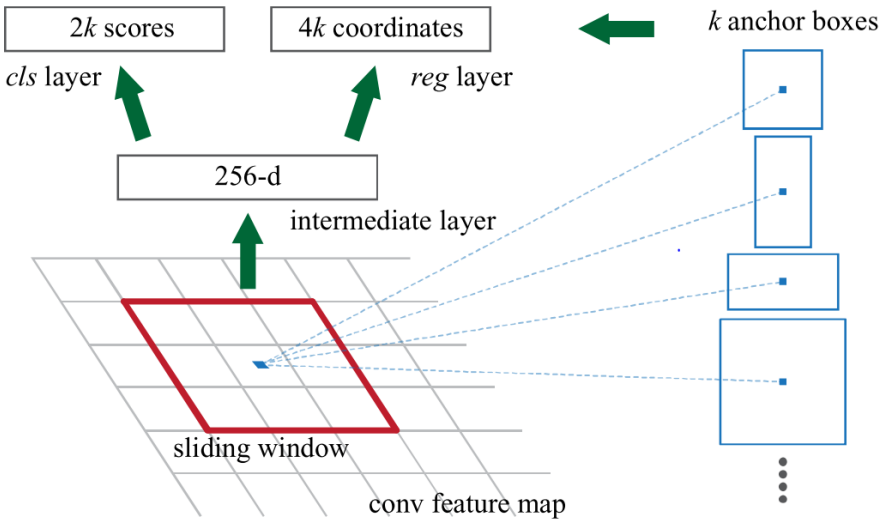


图 2.6 候选区域提取网络[45]

Figure 2.6 Region Proposal Network

锚点框（Anchor）：在每个滑动窗口位置，作者同时预测多个目标候选区域，其中每个位置的最大可能目标候选区域数量表示为k。因为回归层对每个锚点框输出 4 个坐标偏移量，所以共有$4k$个输出，而分类层则输出前景和背景概率，总共有$2k$个分数。通过这些分数可以估计每个候选区域是目标或背景的概率。相对于k个锚点框（Anchor），对k个目标候选区域进行了参数化。锚点框位于相关滑动窗口的中心，并与比例和长宽比相关，如图 2.7。默认情况下，使用 3

个比例和 3 个纵横比，在每个滑动位置产生$k = 9$个锚点框。对于大小为$W \times H$（通常约为 2400）的卷积特征图，总共有 WHk个锚点框。具体锚点框的生成过程如图 2.6 所示。

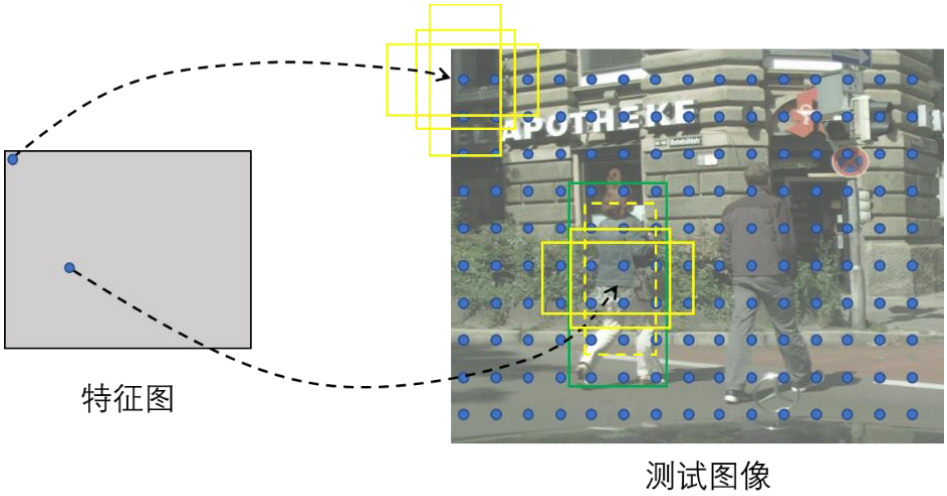


图 2.7 锚点框滑动过程图示

Figure 2.7 Anchor sliding process

在 RPN 训练中，作者为每一个锚点框分配一个二值标签。两种锚点框会被分配到正例标签：（1）和目标标注框（ground-truth）有着最大的交并比（Intersection Over Union，IoU）的锚点框；（2）和任何目标标注框有大于 0.7 的 IoU 的锚点框。一个真实目标可能匹配多个正例的锚点框。通常，第二个条件足以确定正例样本。仍然采用第一个条件的原因是在极少数情况下，第二个条件可能找不到正例样本。如果当前锚点框与所有标注框的 IoU 值均低于 0.3，则将被视作为反例样本。在两个边界之间的锚点框将不会参与计算损失，即作为忽略的锚点框。如图 2.7 中，虚线黄色矩形框则可以被视作正例锚点框。

2.1.1.3 感兴趣区域池化层

感兴趣区域池化层（RoI Pooling Layer）是一个特征采样层，它的输入是图片特征和一组目标候选区域（Proposal），它根据目标候选区域的位置从图片的特征上进行采样，输出固定长度的感兴趣区域特征向量。目的是消除卷积神经网络对输入图像的尺寸要求。卷积神经网络用于分类任务中时，需要输入固定尺寸的图像。当输入图像的尺寸不满足要求时，需要通过裁剪或者缩放等技术进行变换，但是这些操作会造成目标的几何失真和内容缺失，这将降低分类器的精度。

感兴趣区域池化层可以接受任意尺寸的输入，产生相同长度的特征向量，进而消除了网络对输入图像的尺寸要求，有效地解决了这一问题。如图 2.8 展示了感兴趣区域池化层中的量化采样操作过程，它将测试图像中的目标候选框映射到特征图上，然后对该区域进行量化采样，这里使用了4×4的网格。从每个网格中的特征集合中选取特征值最大的值作为输出，最后输出了长度为 16（4×4）的特征向量，该特征向量用于表示目标的特征。

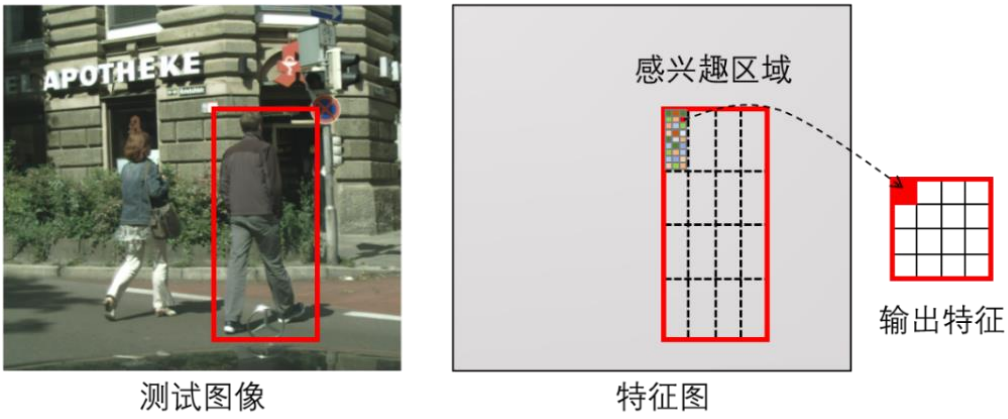


图 2.8 感兴趣区域池化操作

Figure 2.8 RoI pooling operation

2.1.2 单阶段目标检测方法

单阶段检测器因采用较为简单的网络结构，所以对比双阶段检测器有着优秀的检测速度，但是不能达到双阶段检测器的检测精度。单阶段目标检测器的经典代表有 SSD[66]、YOLO[67]和 RetinaNet[68]。我们接下来会介绍到 SSD 和 RetinaNet 的一些细节。

2.1.2.1 SSD

SSD[66]（Single Shot MultiBox Detector）是单阶段目标检测器的经典工作，它是一种基于回归的目标检测器。每一个特征点对应一组锚点框，SSD 为每一个锚点框回归了位置的偏移量并且输出了分类得分。除此之外，SSD 还组合了不同特征层之间的预测结果，这可以处理目标的多尺度问题。它的网络结构极其简单，完全消除了候选区域生成阶段和特征重采样阶段。将所有的计算融入一个单一的网络。

在 2.1.1.2 节中提到的 RPN 可以看作是一种最简单的单阶段检测器，是由于 RPN 使用了单层的特征进行预测，所以没有很好解决多尺度等问题。SSD 使用

了多尺度特征进行检测，具体结构如图所示 2.9 所示。浅层特征具有更加精细的特征，因此有助于目标的定位和小目标的检测；深层特征具有更大的感受野和语义信息，可以提高分类的准确性和帮助大目标的检测。由浅层和深层特征组合类似于特征金字塔，分类和回归的子网在所有层之间参数共享。为了获得更大的感受野，SSD 没有使用 conv5_3 的特征，而是使用更浅的特征 conv4_3 和使用空洞卷积（Dilated Convolution）[69]的 conv7 特征进行检测。

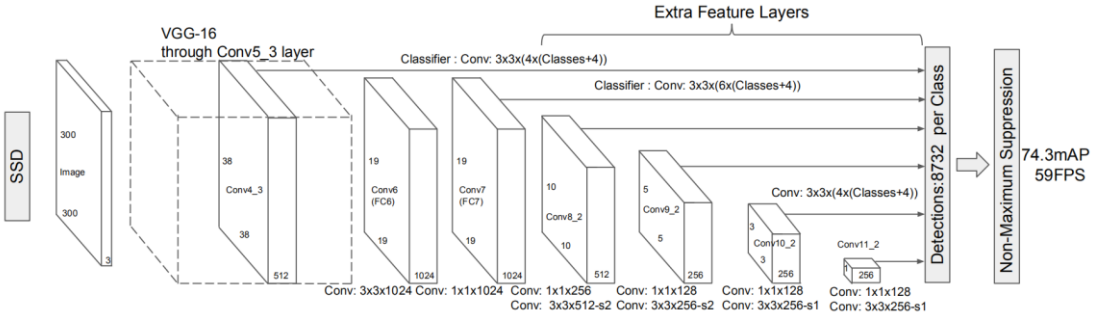


图 2.9 SSD 网络结构示意图[66]

Figure 2.9 Illustration of SSD architecture

2.1.2.2 RetinaNet

RetinaNet[68]是单阶段检测器的杰出代表，具有较高的检测性能和很快的检测速度。作者同时提出 Focal Loss 来解决样本损失不平衡的问题，该损失函数降低了简单样本的学习权重，聚焦于稀疏的难样本的训练，以防止检测器受到大量简单的反例样本的影响。Focal Loss 定义如公式 2.1 所示：

$$FL(p_t) = -\alpha_t(1-p_t)^{\gamma}\log(p_t) \tag{2.1}$$

其中$p \in [0,1]$是对每个类的模型估计概率。为了方便，定义p_t如公式 2.2。

$$p_t = \begin{cases} p & if\ y = 1 \\ 1-p & otherwise \end{cases} \tag{2.2}$$

RetinaNet 网络结构在 ResNet 结构上使用特征金字塔网络[70]（Featrue Pyramid Network, FPN）作为主干网络，生成一个丰富的、多尺度的卷积特征金字塔。特征金字塔网络后端连接两个子网络，其中一个子网络用于进行每个锚点框分类，另一个子网络对每个锚点框进行位置回归。在图 2.10 中，可视化了它的网络结构。RetinaNet 设计简单，结合 Focal Loss 后消除了与双阶段检测器的精度差异，同时运行速度更快。

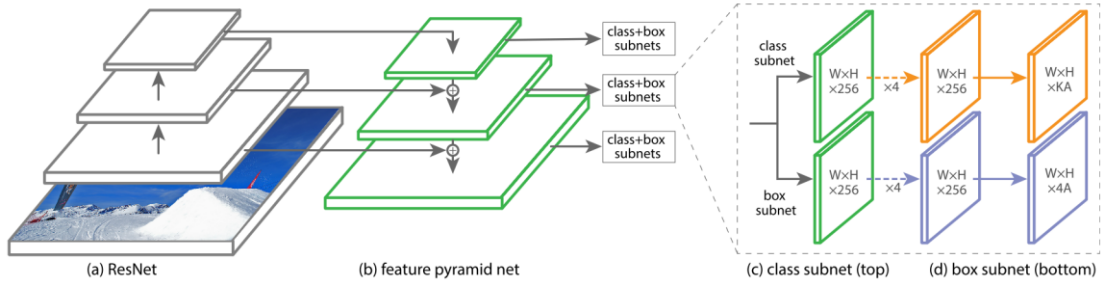


图 2.10 RetinaNet 网络结构示意图[68]

Figure 2.10 Illustration of RetinaNet architecture

2.2 行人检测方法介绍

行人检测是从图像或视频中检测出行人的具体位置，即输出行人位置坐标。计算机视觉任务中的行人检测一般分为 4 个步骤：图片特征提取，候选区域提取，区域特征分类（Object Proposal）和检测窗口融合。早期，候选区域提取的方法是基于穷举策略和底层特征融合，使用最广泛的是滑动窗口遍历法[32]（Sliding Windows）和使用 ACF 检测器[33]产生候选区域。随着深度学习技术的进一步研究，当前候选区域提取的方法也逐渐向深度学习的方法上过渡。

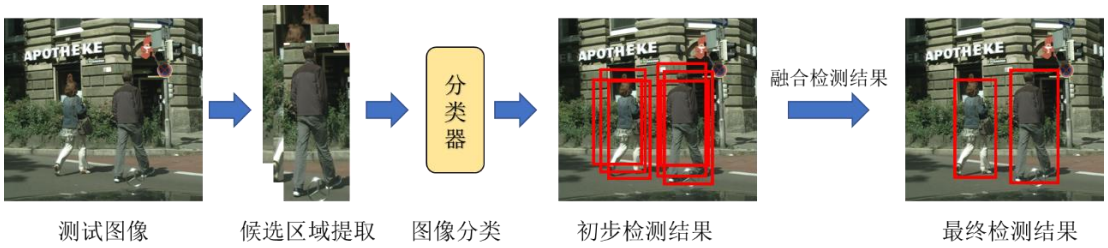


图 2.11 行人检测流程示意图

Figure 2.11 The framework of pedestrian detection

参考论文[46]，我们优化了 Faster R-CNN 训练框架，使其适应行人检测问题。我们使用 ResNet50 替换了 VGG16 作为行人检测的主干网络。ResNet50 比起 VGG16 有着更少的参数量，运行起来速度更快。和 ResNet101 相比，ResNet50 有着合适的感受野，更适合行人目标的尺寸。除此之外，其他的改进还包括：（1）量化了更多的锚点框尺度，覆盖不同尺度的行人目标；（2）忽略训练中的某些训练样本，如骑行的人、坐着的人、人的雕塑和开车的人。与这些目标匹配上的样本，既不算作正例样本也不算做反例样本；（3）使用上采样图片作为输入，

提升检测精度；（4）取消最后一层池化层，以使得最后输出卷积特征在行人目标尺度偏小的情况下，仍具有更高的分辨率。

2.3 遮挡行人检测技术

Occlusion-aware RCNN[53]（OR-CNN）是一种遮挡感知的 R-CNN，这是一种解决遮挡行人检测问题的方法。该方法设计了一个聚合损失函数（Aggregation Loss，AggLoss），强迫目标候选区域相互靠近，并且紧密的定位到相应的目标。除此之外，根据人体先验知识，它把人体分为 5 个部分，然后分别提取 5 个部分的特征，最后使用所有部分的融合特征作为目标特征。它根据人体结构的先验知识可以获得每个区域的局部特征，再结合全局特征做加权求和，最后将融合特征进行分类和回归。如何设计这种加权方式非常重要，为此作者提出一种遮挡处理单元来进行加权，该模块的输出是一个经过 sigmoid 函数得到的值，意味着该行人部分被遮挡的程度，如图 2.12 所示。如果该区域被遮挡，遮挡处理单元则屏蔽该块的特征。因此遮挡处理单元就是一个二分类问题，可以和检测网络联合训练。

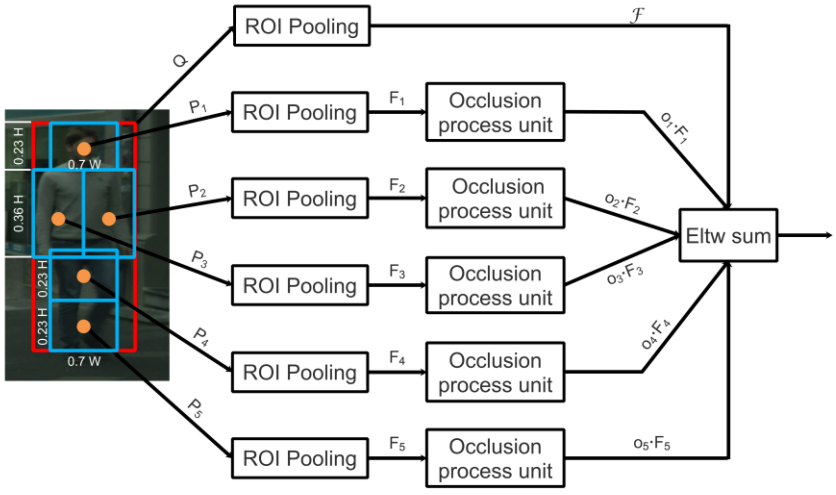


图 2.12 OR-CNN 行人特征提取[53]

Figure 2.12 Part Occlusion-aware RoI pooling unit

针对不同遮挡的行人，Zhang[56]提出 Faster-RCNN+ATT 的方法，以利用注意力机制学习更具有表示性的特征。在卷积网络中采用通道方向上（channel-wise）的注意力来学习遮挡行人特征，FasterRCNN+ATT 方法的流程图如图 2.13 所示。

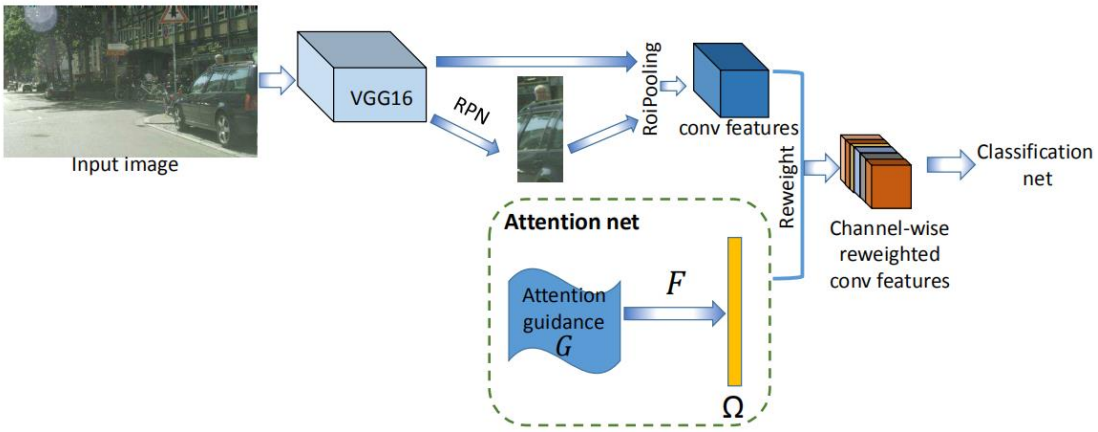


图 2.13 FasterRCNN+ATT 行人检测器流程图[56]

Figure 2.13 Flowchart of attention guided FasterRCNN pedestrian detector

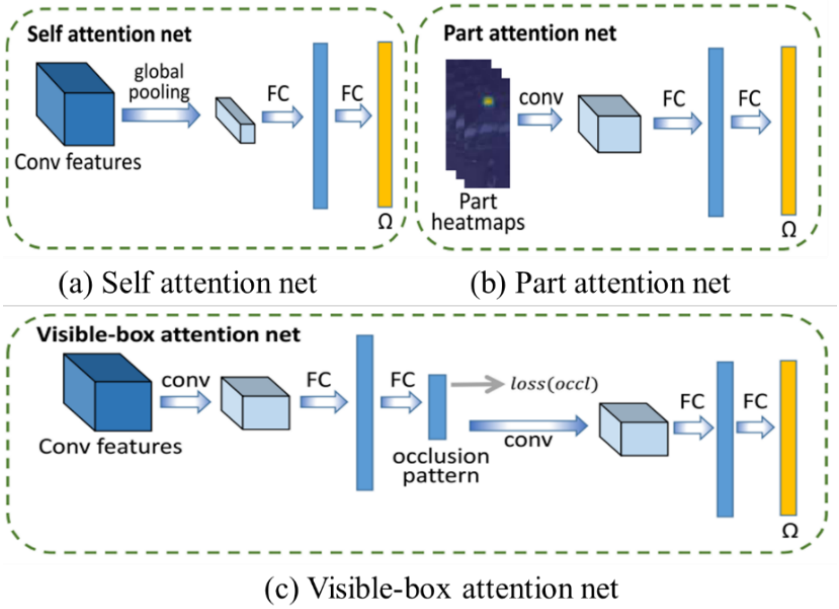


图 2.14 三种不同的注意力网络[56]

Figure 2.14 Three different attention nets use different attention guidances

图 2.14 中显示了作者提出的三种注意力机制引导的网络结构。包括自注意力网络、部件注意力网络和可见框注意力网络，后两种网络使用了额外的可见区域框信息和关键点信息。图 2.14（a）为自注意力网络，此网络和我们后文提出的特征校准网络最为相关，但该自注意力网络隐式地建模了特征和行人之间的关系。而我们所提出的特征校准网络，则是一种显式的建模方。图 2.14（b）中部件注意力网络（Part Attention Net）使用了行人关键点信息。即在训练时需要额外的人体关键点数据集，并且主体网络与该网络是分阶段训练，即不能联合训练。

图 2.14（c）中的可见框注意力网络使用了行人可见框信息，这个网络结构包含多个卷积层和全连接层，增加了一些网络参数。其网络学到的遮挡模式，需经过后端的卷积层和全连接层再次编码，但在编码过程中可能会丢失一些重要信息。

2.4 行人检测数据集

行人检测问题已经被学者们研究了近 20 年。该方向有许多大型的数据集，其中主流的行人检测数据集有 Caltech、CityPersons、KITTI 和 CrowdHuman。

（1）数据集：Caltech

Caltech[72]数据集包含 10 个小时的街景视频，由车辆上的单目摄像头所采集。这个数据集的最大挑战是包含大量低分辨率和遮挡的行人样本。其中从 set00 到 set05 中采样出的 42782 幅图像用于训练，从 set06 到 set10 集合中采样的 4024 幅图像用于测试。

（2）数据集：CityPersons

CityPerson[46]数据集是建立在语义分割数据集 Cityscapes[71]上的。该数据集包含德国的 18 个城市、三个不同季节和各种天气情况。总共有 5000 幅图像，其中 2975 幅图像用于训练，500 幅图像用于验证，剩下 1525 幅图像用于测试。图片均为高分辨率图像，具有2014×2048分辨率。这个数据集包含更多的拥挤场景，同时也包含更多的严重遮挡的行人样本。

（3）数据集：KITTI

KITTI[73]数据集由 7481 幅训练图像和 7518 幅测试图像组成，包括大约 8 万个汽车、行人和骑行人的标注。KITTI 使用三个 PASCAL[74]风格的指标评估平均精确度（mAP）：简单，中等和困难。困难等级根据行人高度、遮挡率和被截断情况划分。

（4）数据集：CrowdHuman

CrowdHuman[75]是一个比较新的人体检测数据集，由 Shao 使用 150 个关键字在 Google 图像搜索引擎中进行图像查询所得。该数据收集到的图片覆盖了全球 40 多个不同的城市，包括各种各样的活动，例如聚会、旅行和体育等。该数据集还包括多种视角，例如监控视角和水平视角等。为了平衡数据，每个关键字限制最大抓取 500 幅图像。该数据集总共包含约 2.5 万幅图像，其中 15000 幅图像为训练集，4370 幅图像为验证集，5000 幅图像为测试集。

本文主要使用经典的 Caltech 行人数据集和使用最广泛的 CityPersons 数据集。

2.5 行人检测性能评价

如果一个检测框（BBdt）和一个标注框（BBgt）的重叠面积满足一定数值，那么它们会成为一对潜在的匹配。行人检测评估参考了 PASCAL[74]的评测标准，即重叠面积超过 50%，则认为这个检测框匹配到了标注框：

$$a_{\mathrm{o}} = \frac{area(BB_{dt} \cap BB_{gt})}{area(BB_{dt} \cup BB_{gt})} > 0.5 \qquad (2.1)$$

除此之外，还需要注意每个 BBdt 和 BBgt 最多只能匹配一次。通过一个贪婪算法来执行匹配，以防止出现的任何歧义。首先，将检测结果根据其置信度进行降序排序，最大置信度的检测结果会被匹配；如果一个检测框匹配到了标注框，那么具有最大重叠面积的匹配将会被保留。在极少数的情况下，这个分配结果是次优的，例如拥挤场景，但是这个算法在实际中仅受到了较小的影响。没有匹配到的 BBdt 算作假正例（False Positives，FP），没有匹配到的 BBgt 是假反例（False Negatives，FN）。

表 2.1 分类结果混淆矩阵

Table 2.1 Classification result confusion matrix.

真实情况	预测结果	
	正例	反例
正例	TP（真正例）	FN（假反例）
反例	FP（假正例）	TN（真反例）

通常使用丢失率和单位图片下的虚警率（False Positives Per Images, FPPI）的数量曲线来比较不同行人检测方法的性能。这些指标对于一些汽车辅助系统和自动驾驶等任务是合理的，因为通常这些任务会要求给定单位图片虚警率的上限，并且要求评测标准与行人密度无关。为了反应丢失检测的情况，可使用对数平均丢失率（Log-average Miss Rate, MR^{-2}）来评测检测器的性能，它在对数空间的 10^{-2} 到 10^{-0} 范围内 9 个 FPPI 的平均值，简称为 MR。

在评测时可根据目标的可见率，将数据集分为几个子集[14][56][53][59]。分别为 All、Reasonable、Partial、Heavy、None。这样的划分方便观测算法在不同分辨率和不同遮挡程度上的性能。Caltech 数据集的划分情况如表 2.2 所示，而

CityPersons 数据集的划分情况如表 2.3 所示。需要注意的是他们的 Partial 集合上定义稍有不同。

表 2.2 Caltech 数据集子集划分

Table 2.2 Caltech dataset subset division.

子集名称	可见率
All	[20%, +∞]
Reasonable	[65%, +∞]
Partial	[65%, 100%]
Heavy	[20%, 65%]
None	[+∞, +∞]

表 2.3 CityPersons 数据集子集划分

Table 2.3 CityPersons dataset subset division.

子集名称	可见率
All	[20%, +∞]
Reasonable	[65%, +∞]
Partial	[65%, 90%]
Heavy	[20%, 65%]
None	[+∞, +∞]

2.6 本章小结

本章详细地描述了通用目标检测器、行人检测器和遮挡行人检测器。同时，本章介绍了经典的目标检测器，包括双阶段检测器和单阶段检测器。双阶段的检测模型有 Faster R-CNN，单阶段的检测模型有 SSD 和 RetinaNet。综述了几种经典的遮挡行人检测问题的方法并介绍了几种数据集。

第3章　基于循环特征增强的行人检测

为了解决遮挡及小目标尺度行人特征提取问题，本文提出了一种自适应特征增强的深度网络，称之为环状循环网络。为了更好地发挥环状循环网络的优势，我们结合行人实例分解策略，使得行人检测性能有了进一步的提升。环状循环网络具有特征适配能力，可根据行人的遮挡程度和尺寸，自动生成对应的自适应特征。行人分解策略有效地解决了遮挡行人特征和非遮挡行人特征之间差异的问题，更好地发挥了环状循环网络的能力。图 3.1 展示了特征金字塔网络[70]（Feature Pyramid Network, FPN）结构到环状循环网络（CircleNet）结构的变化。图 3.1（a）是特征金字塔网络，它在全卷积网络的基础上增加了一条由深层到浅层的通路，使得深层的语义信息可以回传给浅层；图 3.1（b）是 PANet[76]（Path Aggregation Network）网络结构，它在 FPN 的基础上增加了一条自底向上的网络分支，这使浅层信息可以通过更短的通路到达深层；图 3.1（c）是本文中提出的环状循环网络，它多次使用自底向上和自上而下的两个分支，更好地融合了深层和浅层的特征，并且促进深层和浅层之间的信息交换。这样的循环使得网络更具有特征适配的能力，可针对不同外表的样本生成具有针对性的特征表示。

3.1 环状循环网络

对于困难目标的识别，我们人类也很难看一眼就识别出来，所以往往需要多次观察并专注于图像的特定区域。受此启发，我们提出了一种特征学习框架 CircleNet，如图 3.1（c）所示。该框架主要包含三大优势：

（1）往复融合特征。CircleNet 通过模仿人类在观察低分辨率和被遮挡的物体时如首次无法明确识别物体，就会进行多次识别的方法。采用了环状循环结构来实现特征自适应，环形结构由自上而下的分支和自下而上的分支构成，它们以往复的方式融合特征，这意味着各种外观的行人在不同的循环中被处理。更有助于学习自适应特征表示。

（2）提升遮挡和低分辨率目标的特征表示。模仿人类专注于图像特定区域的行为，CircleNet 通过多个自上而下和自下而上的特征融合路径来增强特征适应性，从而同时提升遮挡和低分辨率目标的特征表示。但当吸收来自更多通路的

信息时，更多参数可能产生过拟合，因此我们需要注意这些附加层的模型参数容量。我们提出对这些附加层使用权重共享策略，并通过实验验证共享不同循环之间自上而下和自下而上的路径的网络参数的可行性，以保持模型能力和泛化能力之间的平衡。

（3）通用性更强。CircleNet 是一种通用网络体系结构。FPN 和 PANet 可以看作是 CircleNet 的特例，相比于这两种网络 CircleNet 更具有通用性。

在本节中，我们首先介绍 CircleNet 的具体结构，然后解释特征适配含义。随后提出行人实例分解训练策略并进行了网络优化，最后进行了实验测评。

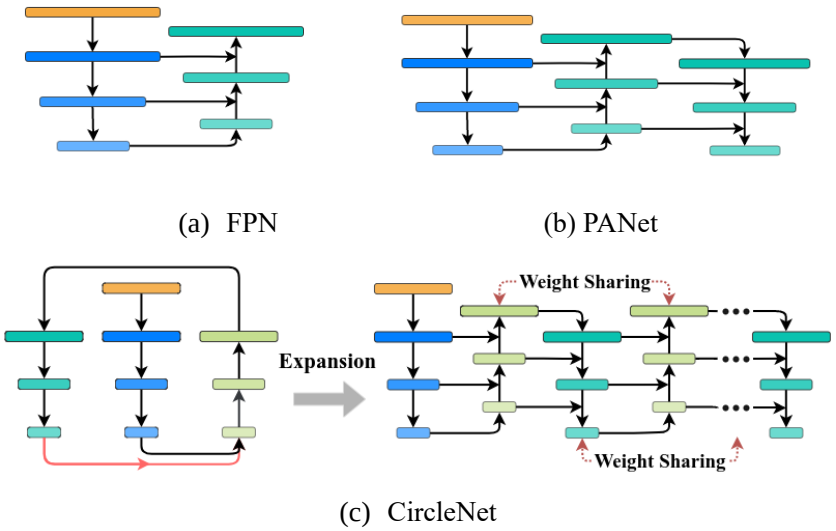


图 3.1 从特征金字塔网络（FPN）到环状循环网络（CircleNet）的过程。（a）特征金字塔网络；（b）路径汇聚网络；（c）本文提出的特征增强循环网络，它通过使用权重共享的一组金字塔结构所实现。

Figure 3.1 From Feature Pyramid Network to CircleNet. (a) Feature Pyramid Network (FPN). (b) Path Aggregation Network (PANet). (c) Our proposed CircleNet is implemented as a set of feature pyramids using weight sharing path augmentation.

3.1.1 环状循环网络结构

对于自上而下的通路（基数分支），我们用o_n表示第n个特征层，这里$n = 1,\ldots,N$，o_{n+1}和e_n分别表示输入和侧输出。e_n存储了来自当前层的信息，并且o_{n+1}携带了来自深层的高层语义信息。相似地，对于自下而上的通路（偶数分支），我们用e_{n+1}表示第$(n+1)$个特征层，并且e_n和o_{n+1}分别表示输入和侧输出。o_{n+1}存储了来自当前层的信息，e_n被用于产生新的特征。同时级联多个自上而下和自

下而上的分支可以实现一个级联的特征金字塔网络（Cascaded Feature Pyramid Network）。

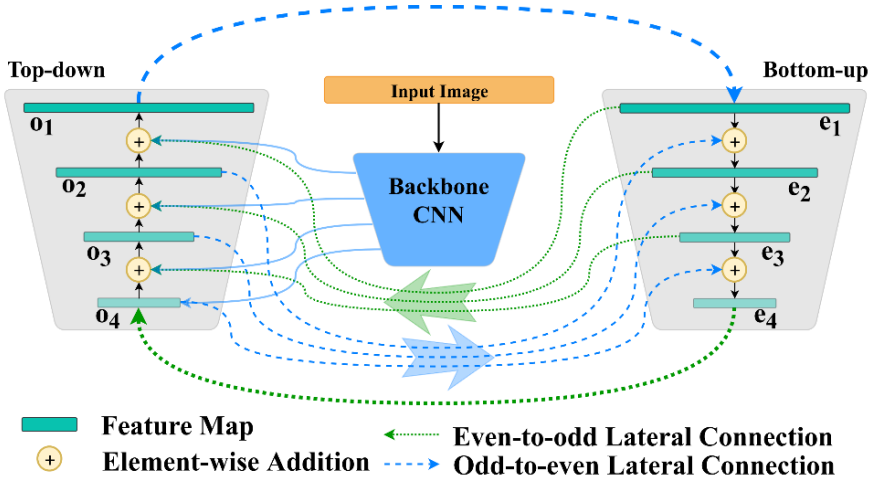


图 3.2 CircleNet 网络结构图。自上而下的分支通过上采样和级联深层特征加强了语义信息；自下而上的分支通过下采样和级联来自浅层的特征增大了感受野，同时聚集了上下文信息。

Figure 3.2 CircleNet architecture. A top-down branches enforce semantics by up-sampling and concatenating deep-layer features. A bottom-up branches enlarge the receptive field and incorporate context information by down-sampling and concatenating features from shallow layers.

为了升级这个级联的网络到一个环状循环网络，我们在不同循环之间共享了卷积参数，如图 3.1（c）所示。通过这种方式，我们以少量的学习参数学习到了适应不同行人的特征。这就意味着我们只需要学习一组奇数分支和一组偶数分支的权重$\{w^e, w^o\}$，然后通过权值共享，更新这个级联结构到整个环状循环网络，使得每一次循环的过程中，都可以学习到不同的特征：

$$\boldsymbol{x}_t = F(\boldsymbol{x}_{t-1}, w^e, w^o) \tag{3.1}$$

其中$\boldsymbol{x}_t = \{o_n^t, e_n^t\}$、$w^e = \{w_n^e\}$、$w^o = \{w_n^o\}$、$n = 1, \ldots, N$、$t = 1, \ldots, T$。$n$是特征层的序号，$t$是循环的序号，$F(\cdot)$表示特征提取函数。

3.1.2 往复式的特征自适应

环状结构通过一种往复式地连接多个自上而下的通路和自下而上的通路的方式实现了特征融合和特征适应。每一个通路使用侧连接融合来自主干网或者其他通路的特征。针对自上而下的通路设计，我们遵循 FPN 的实现方式。在任意一个循环过程中，即第t个循环时：

$$o_n^t = F_n(e_n^t, o_{n+1}^t, w_n^e) = w_n^{e_{11}} * \left(w_n^{e_{33}} * e_n^t\right) + \uparrow o_{n+1}^t \quad (3.2)$$

这里$w_n^e = \{w_n^{e_{11}}, w_n^{e_{33}}\}$是$1 \times 1$和$3 \times 3$卷积滤波器，用于融合特征生成$o_n^t$。↑表示上采样操作，该操作由最近邻插值实现。

相似地，e_{n+1}^t是第$(n+1)$个自下而上层的特征，它是通过融合来自自下而上层的特征o_{n+1}^{t-1}和它前一个自下而上特征层e_n^t，如下：

$$e_{n+1}^t = F_n(e_n^t, o_{n+1}^{t-1}, w_{n+1}^o) = w_{n+1}^{o_{11}} * \left(w_{n+1}^{o_{33}} * o_{n+1}^{t-1}\right) + \downarrow e_n^t \quad (3.3)$$

这里$w_n^o = \{w_n^{o_{11}}, w_n^{o_{33}}\}$是$1 \times 1$和$3 \times 3$卷积滤波器，用于融合特征和生成$e_{n+1}$。↓表示下采样操作，该操作由一个步长为 2 的$3 \times 3$卷积实现。

在前传过程中，自上而下的通路通过上采样高层特征生成更高分辨率的特征，该高层特征是来自高层金字塔的特征，其缺乏精确位置信息但是具有强语义。然后，通过横向连接，使用来自主干网或自下而上的路径的通路特征来增强这些特征。每个横向连接将自下而上的通路和自上而下的通路合并成相同空间尺寸的特征图。自下而上的通路通过降低特征分辨率但增强其语义进一步融合了特征。降低特征图的分辨率可以扩大感受野并提取上下文信息，这对于检测低分辨率和被遮挡的行人至关重要。

根据公式 3.2 和公式 3.3，通过多次使用自上而下的通路和自下而上的通路，CircleNet 实现了往复式的特征融合，从而同时提供了具有强语义和上下文信息的特征。

3.2 行人实例分解

通过往复式的特征融合，这个环状网络有着潜在学习适应性特征的能力，这些特征可以适应不同外表的行人。我们经验地把这些不同外表的行人分配到不同循环中，也就是说使用不同循环的特征表示不同外表的行人。因此，普通的样本和难样本可以使用合适的特征进行处理。在实验中，我们观测到深层循环更加关注遮挡行人样本，然而没有遮挡的样本在浅层循环上检测的更好。因此，我们提出了行人实例分解策略，该策略根据训练损失或者遮挡率把行人样本分解到不同循环中。同时根据分辨率把行人样本分解到不同深浅特征层上训练。

$$D = \bigcup_t D^t, D^t \cap D^{(t+1)} \neq \emptyset \quad (3.4)$$

这里$D^t = \sum_n D_n^t$。D^t是一个行人实例的子集，在这个集中存放简单的样本。$D^{(t+1)}$是一个更大的子集，在这个集合中包含一些比D^t更难一些的样本。这些难样本可以通过他们的遮挡率或者训练损失来定义。

在每一个循环中，沿着深浅特征层方向，D^t又被分解成几个相互不重叠的子集，如下：

$$D^t = \bigcup_n D_n^t, D_n^t \cap D_{n+1}^t = \emptyset \tag{3.5}$$

这里D_{n+1}^t是行人实例的一个子集，在这个子集中样本的分辨率高于D_n^t。这里我们继承了 FPN 的样本分解策略，我们分配高分辨率的样本到深层，同时分配低分辨率的样本到浅层。

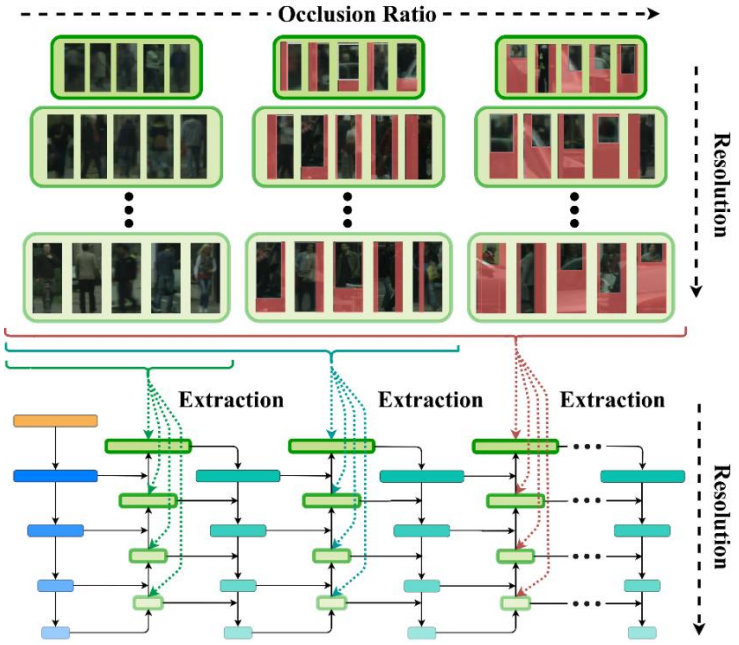


图 3.3 行人实例分解训练策略示意图

Figure 3.3 An illustration of instance decomposition during the training phase.

3.3 网络优化

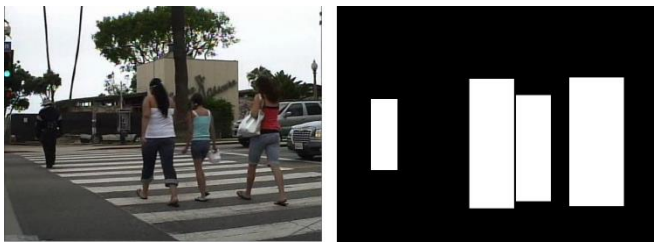


图 3.4 行人注意力区域分割标注

Figure 3.4 Pedestrian attention segmentation annotation

为了使网络更加关注行人区域，我们为行人检测网络添加了一个分割分支。即使用基于弱监督行人框的语义分割[79]训练方法。行人相对于图片的尺度较小，在下采样的特征层上它的精细掩膜类似于矩形[80]，所以我们将标注矩形框转化为弱标注的分割掩膜，如图 3.4 所示。最后将该掩膜作为监督信息，用于训练分割分支。由于自上而下的网络通路汇聚了来自深层网络的语义信息，我们因此添加了一个伪分割损失在自上而下的网络通路上。通过引入分割层，使得特征关注行人区域，并且抑制来自聚类背景上的反例样本。

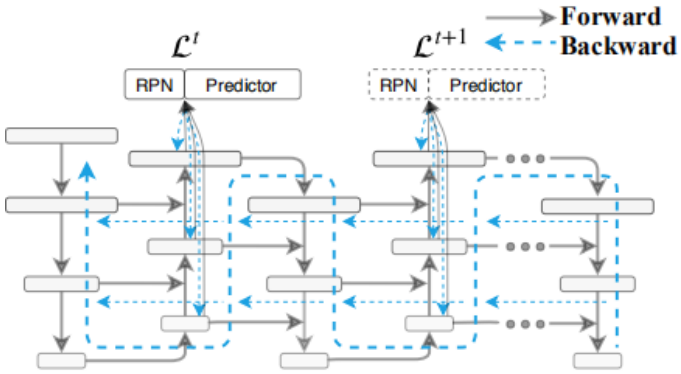


图 3.5 CircleNet 网络优化过程。在学习过程中，CircleNet 上的前向连接（带箭头的实线）用于特征提取和分数预测；反向连接（带箭头的虚线）用于梯度传播。

Figure 3.3 CircleNet optimization process. During the learning procedure, forward connections (solid lines with arrows) along the CircleNet are for feature extraction and object score prediction. Backward connections (dashed lines with arrows) are for gradient propagation.

我们通过多个 RPN 和预测器（Predictor）同时优化分类和回归任务来训练这个网络。交叉熵损失函数被用分类任务，mooth_{L1}[43]损失函数用于回归任务。损失函数定义如下：

$$\mathcal{L} = \sum_t \mathcal{L}^t = \sum_t \left(\sum_n \mathcal{L}_{rpn\_cls}{}_n^t + \mathcal{L}_{rpn\_reg}{}_n^t + \mathcal{L}_{cls}{}_n^t + \mathcal{L}_{reg}{}_n^t \right) + \mathcal{L}_{seg}{}^t \quad (3.6)$$

$\mathcal{L}_{rpn\_cls}$ 和 $\mathcal{L}_{cls}$ 用于优化目标候选区域和最终检测框的分类，回归损失 $\mathcal{L}_{rpn\_reg}$ 用于优化目标候选区域和标注框之间的位置偏移，$\mathcal{L}_{reg}$ 用于优化检测结果和标注框之间的位置偏移。$\mathcal{L}_{seg}$ 是辅助的分割损失。网络前传和反传的过程如图 3.5 所示，它显示了特征和梯度可以沿着循环被传递。$\mathcal{L}_{cls}$ 使用的是二值交叉熵损失函数，具体定义如公式 3.7 所示：

$$\mathcal{L}_{cls} = -[y_i \cdot \log p + (1 - y_i) \cdot \log(1 - p)] \quad (3.7)$$

这里y_i是目标的类别标签，p是分类器预测的类别概率。$\mathcal{L}_{seg}$也采用二值交叉熵损失函数，不同的地方在于y_i表示像素的标签，p代表对每个像素的预测概率。

$\mathcal{L}_{reg}$使用了smooth_{L1}损失函数，具体形式如公式 3.8 所示：

$$\mathcal{L}_{reg}(t,v)=\sum_{i\in x,y,w,h}\text{smooth}_{L1}(t_i-v_i) \tag{3.8}$$

其中$v=(v_x,v_y,v_w,v_h)$为标注框，$t=(t_x,t_y,t_w,t_h)$为网络预测的检测框。这里使用到的smooth_{L1}损失函数定义如下：

$$\text{smooth}_{L1}(x)=\begin{cases}0.5x^2 & \text{if } |x|<0.5\\ |x|-0.5 & \text{otherwise}\end{cases} \tag{3.9}$$

$\mathcal{L}_{rpn_cls}$的形式和$\mathcal{L}_{cls}$相同，均采用二值交叉熵损失函数；$\mathcal{L}_{rpn_reg}$的形式和$\mathcal{L}_{reg}$相同，均采用smooth_{L1}损失函数。

从学习的角度来看，CircleNet 实现了一种特殊的分类器集成。对于多个循环上的多个特征层，我们有多个分类器，每个分类器负责行人实例的一个子集。在深度学习框架中，难样本具有很大的训练损失，这意味着它们在学习过程中具有较大的权重，这实际上是一种促进学习的策略。在训练阶段，使用多个基本分类器处理外观不同的行人实例，例如分辨率和遮挡率。在测试阶段，最大分类分数用于确定最终检测结果。

通过基于特征往复的融合过程，我们实际上可以增强在不同子集上学习的基本分类器，最后在困难样本上获得的特征也可以适应普通样本。这不仅有利于行人检测，而且在深度学习框架中带来了一种新的集成方法。

3.4 实验结果与分析

为了验证本章所提出的环状循环网络的有效性，本文使用 ResNet50 作为实验的基础网络，在较为常用的 Caltech 和 CityPersons 数据集中测试了本章提出的方法。我们进行消融实验以验证 CircleNet 在复杂的交通场景中对行人检测的有效性。后面几个小节将分别介绍实验设定、实验分析以及和相关方法的结果对比。

3.4.1 实验设定

我们使用带有 FPN 结构的 ResNet50 作为主干网。为了与其他方法公平的对比，将 Caltech 数据集中的图像分辨率上采样到900×1200，并且使用 CityPersons 数据集进行预训练。在 CityPersons 数据集上评测时，我们使用了[46]的设定，上采样图像分辨率采用原始分辨率的 1.3 倍，并且使用 ImageNet 预训练的主干网。

我们使用 1 块 K80 GPU 进行网络的训练。对于 Caltech 数据集，模型总共训练了 35k 次迭代。初始学习率设定为 0.002，在 25k 次训练迭代后，学习率衰减 10 倍，并且使用随机梯度下降算法（Stochastic Gradient Descent，SGD）进行优化，每个 GPU 上的小批量包含 2 幅图像。权重衰减设定为 0.0001，动量设定为 0.9。对于 CityPersons 数据集，模型总共训练了 80k 次迭代。初始学习率设定为 0.0025，在 50k 次训练迭代后，学习率衰减 10 倍，每个 GPU 上的小批量包含 1 幅图像。评测数据集分别使用 Caltech 数据集和 CityPersons 数据集，详细介绍见 2.4 节。评测标准使用对数平均丢失率（MR^{-2}），详细介绍见 2.5 节。

3.4.2 环状循环网络循环次数选择实验

表 3.1 FPN 和 CircleNet 在 Caltech 测试集上的性能对比（Height ≥ 50）。MR^{-2} 作为评测指标，分数越低代表性能越好。

Table 3.1 Detection performance of FPN and CircleNet on the Caltech test set （Height ≥ 50）. MR^{-2} is used to compare the performance. Lower score indicates better performance.

Model	Description	All	None	Partial	Heavy	Reasonable
FPN[70]	Baseline	32.21	12.72	37.40	82.96	15.79
CirCleNet-1/2	PANet structure	30.02	12.76	37.82	79.20	15.80
CirCleNet-1	One circle (T=1)	28.74	**10.85**	**34.28**	79.87	**13.75**
CirCleNet-2	One circle (T=2)	**28.12**	11.58	36.49	75.08	14.63
CirCleNet-3	One circle (T=3)	28.69	11.66	36.32	**75.01**	14.72
FPN+	Baseline	25.88	12.58	32.03	64.01	14.85
CirCleNet-2+	Two circles	**24.14**	12.84	**28.74**	**54.66**	15.02

表 3.2 FPN 和 CircleNet 在 Caltech 测试集上的性能对比（Height ≥ 20）

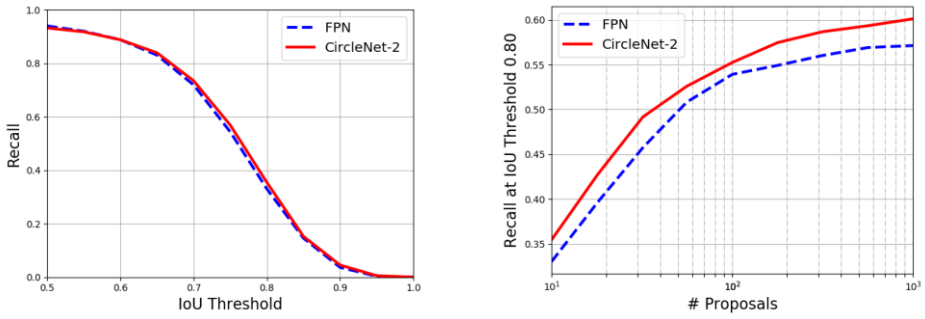
Table 3.2 FPN and CircleNet detection performance on the Caltech test set （Height ≥ 20）.

Model	Description	All	None	Partial	Heavy
FPN[70]	Baseline	59.85	48.39	69.41	89.71
CirCleNet-1/2	PANet structure	57.26	45.48	68.98	88.72
CirCleNet-1	One circle (T=1)	**57.00**	**45.21**	68.29	88.86
CirCleNet-2	One circle (T=2)	57.25	45.62	69.27	**88.19**
CirCleNet-3	One circle (T=3)	58.71	48.89	**68.18**	88.65
FPN+	Baseline	57.02	48.51	64.33	79.33
CirCleNet-2+	Two circles	**55.36**	**46.97**	65.77	**75.05**

我们定义信息通过一次自底向上通路和一次自顶向下通路为一次循环。对比中使用 FPN 作为基准网络。CircleNet-1/2 是半个循环结构，它增加一条自底向上的通路，将 FPN 扩展为 PANet[76]。有 1 次、2 次和 3 次的循环结构在实验中分别被叫做 CircleNet-1、CircleNet-2 和 CircleNet-3。结果表明到 CircleNet-3 已经可以

针对不同的行人实例和难样本提供足够的特征层和特征表示，因此在实验中没有继续增加更多的循环次数。

表 3.1 和表 3.2 分别为$Height \geq 50$和$Height \geq 20$样本集合的评测性能。从表格 3.1（$Height \geq 50$）可以看出，在“All”集合上，CircleNet-1 的检测结果优于 FPN 和 CircleNet-1/2，同时 CircleNet-2 又优于 CircleNet-1 网络 0.62%。CircleNet-3 的性能略低于 CircleNet-2，这可能是由于增加循环次数会增加网络的训练难度引起的。在表格 3.2（$Height \geq 20$）中，也呈现相似的趋势，在“All”集合上，CircleNet-2 和 CircleNet-1 性能仅相差 0.25%。在$Height \geq 50$ 的“All”和“Reasonable”子集上，CircleNet-2 的对数丢失率分别好于 FPN 方法 3.09%（28.12% vs. 31.21%）和 1.16%（14.63% vs. 15.79%）。在$Height \geq 50$ 的“None”和“Partial”遮挡子集上，CircleNet-2 性能优于 FPN 方法 1.14%（11.58% vs. 12.72%）和 0.91%（36.49% vs. 37.40%）。在$Height \geq 50$的“Heavy”遮挡子集上，CircleNet 的性能高于 FPN 方法 7.84%（75.08% vs. 82.96%）。从表 3.2 可以看出，在$Height \geq 20$的“All”子集上，CircleNet-2 优于 FPN 方法 2.60%（57.25% vs. 59.85%），并且优于 PANet（CircleNet-1/2）方法 1.9%（28.12% vs. 30.02%）。为了公平的对比，用于比较的 PANet 没有使用适应性特征池化层（Adaptive Feature Pooling），我们只是比较主干网的结构。如表 3.1 和表 3.2 的最后两行所示，当使用严重遮挡样本训练网络时（称之为 FPN+和 CirCleNet-2+），CircleNet 仍然优于基准方法，特别是对于低分辨率的样本。



(a) 不同 IoU 阈值下的召回率　　(b) 不同候选区域数量下的召回率

图 3.6 环状循环结构对于 RPN 的影响。

Figure 3.6 Evaluation of the effect of CircleNet for region proposal network (RPN). (a) Recall rates under different IoU thresholds. (b) Recall rates under different numbers of region proposals.

同时提出的 CircleNet 也提高了候选区域提取网络（RPN）的性能。我们使用了候选区域评测的两个常用指标进行评测。第一个指标为不同 IoU 阈值下的召回率，第二个指标为在不同数量的候选区域下召回率的变化。如图 3.6 所示，在这两个指标方面，CircleNet 的表现均优于基准方法 FPN。尤其是在图 3.6（b）中，固定目标候选区域的数量，阈值为 0.8 的召回率有明显提升。这说明环状循环网络不仅可以提升检测结果，对候选区域提取网络也有积极的影响。

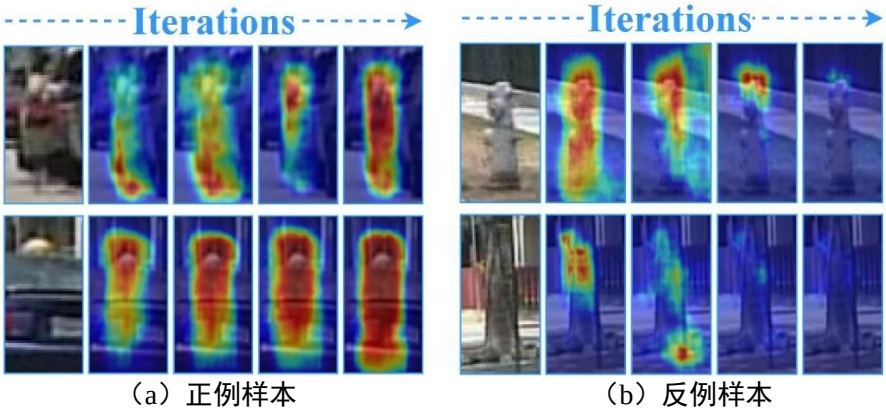
表 3.3 在 Caltech 数据集上 CircleNet 的检测速度对比

Table 3.3 Comparison of the detection speeds on Caltech dataset.

Method	FPN	CircleNet-1	CircleNet-2	CircleNet-3
Inference Time (ms/image)	51	61	74	87

在表 3.3 中，我们比较了 FPN 和 CircleNet 的测试速度。处理一幅图片，FPN 需要 51ms，CirleNet-1、CircleNet-2 和 CirleNet-3 分别需要 61ms、74ms 和 87ms。这表明 CircleNet 由于增加了计算损耗，检测速度有一定下降。但这种计算损耗的适量增加却能明显提升检测精度，如表 3.1 和表 3.2 所示。而在某些特定场景中，如无人驾驶技术中，为保障行人及驾驶员人身安全，精准检测往往更为重要。因此，在综合在考虑检测精度和运行速度的平衡之后，我们选择 CircleNet-2 作为后续的实验基准。CircleNet-2 在 Tesla 1080TI GPU 平台上的运行速度为 13.5FPS，完全可以满足一些实际的应用需求。

3.4.3 行人语义分割分析实验



（a）正例样本　　（b）反例样本

图 3.7 训练迭代过程中行人的分割结果变化。正样本的特征得到了增强，而反例样本的特征得到了抑制。

Figure 3.7 With segmentation loss, the feature representation for the positive examples are enforced while that for the negative examples are depressed.

在 3.3 节中，我们提出使用弱监督的行人语义分割框作为辅助监督。通过该辅助分割损失，激活了 CircleNet 行人边界上的特征。图 3.7（a）展示出了对于正例样本，在训练过程中，激活掩模区域逐渐填充目标的边界框。这可以让网络更加关注于行人区域。图 3.7（a）中第 2 列展示了一个被遮挡的行人，它激活的区域很好的保留了上下文信息区域。与此同时，弱监督的行人语义分割作为像素分类任务，也有助于抑制来自复杂背景的负样本。如图 3.7（b）所示，第 1 列的消火栓和第 2 列的树木的激活区域都被逐渐抑制。

图 3.8 中展示了 6 幅网络辅助分割分支的最终输出结果，这些输出的特征经过了 Sigmoid 函数。分割网络高亮了行人区域，使得主干网络更加关注该行人区域，抑制了背景的干扰。同时激活区域保留了行人的上下文信息区域。

图 3.8 弱监督行人语义分割结果

Figure 3.8 Weakly supervised pedestrian semantic segmentation results

3.4.4 行人实例分解验证实验

表 3.4 CircleNet 行人实例分解和分割损失性能对比（Height ≥ 50）

Table 3.4 Detection performance of CircleNet with instance decomposition and segmentation loss on the Caltech test set（Height ≥ 50）.

Model	Description	All	None	Partial	Heavy	Reasonable
CirCleNet	w/o	24.14	12.84	28.74	54.66	15.02
CirCleNet+ID1	None	22.86	12.33	27.68	50.40	14.54
CirCleNet+ID2	By loss (OHEM)	22.37	12.92	28.45	48.35	14.83
CirCleNet+ID3	All-to-hard	23.84	12.25	28.00	55.19	14.48
CirCleNet+ID4	Easy-to-hard	**21.62**	**11.83**	**26.45**	**48.70**	**13.78**
CirCleNet+Seg	Segmentation	21.57	12.44	27.50	45.47	14.38
CirCleNet+	ID4+Seg	**18.05**	**8.42**	**20.27**	**44.53**	**10.21**

表 3.5 CircleNet 行人实例分解和分割损失性能对比（Height ≥ 20）

Table 3.5 Detection performance of CircleNet with instance decomposition and segmentation loss on the Caltech test set（ Height ≥ 20）.

Model	Description	All	None	Partial	Heavy
CirCleNet	w/o	55.36	46.97	65.77	75.05
CirCleNet+ID1	None	54.54	47.42	66.11	69.42
CirCleNet+ID2	By loss (OHEM)	55.55	47.15	65.31	**72.25**
CirCleNet+ID3	All-to-hard	55.05	46.40	65.65	76.48
CirCleNet+ID4	Easy-to-hard	**52.69**	**44.29**	**64.83**	73.09
CirCleNet+Seg	Segmentation	54.62	47.12	66.29	71.55
CirCleNet+	ID4+Seg	**46.42**	**37.48**	**59.26**	**66.28**

在实验中，我们对比了多种行人实例分解策略。在表 3.4 和表 3.5 中，分析了这些策略下的性能。“None”表示没有使用任何分解，在不同循环中，所有的样本都被使用；“By loss”表示使用损失函数的数值进行分解，将损失比较大的样本输入到深层循环中训练，反之则输入到浅层循环中训练。归一化并且缩放这些样本的损失值到一个固定的范围。最后定义样本的学习权重为：$w = \frac{l-lmin}{lmax-lmin} \times (1-\alpha) + \alpha$；这里 l 是训练样本是损失值，我们经验性地设定$\alpha = 0.7$。“All-to-hard”表示全部样本先都被送入浅层循环中训练，再将严重遮挡的样本送入深层循环中训练；“Easy-to-hard”表示将非遮挡样本和一般遮挡的样本都送入浅层循环中训练，再将严重遮挡的样本送入深层循环中训练；“Segmentation”表示使用弱监督的行人语义分割作为额外监督信息。在以上分解策略中，“Easy-to-hard”取得了较好的性能（52.69%）。这是因为容易样本包含更多的有用信息，将其输入到浅层循环中，网络不容易丢失有用的信息，并且深层循环可以提供更多的上下文信息给严重遮挡的行人。在“Easy-to-hard”基础上使用辅助语义分割损失的方法，我们命名为 CircleNet+，可以看出通过使用语义分割性能得到了进一步提升，在“All”集合上降低至 46.42%。

为了更好地理解环状循环网络如何提升了检测性能，我们对 CircleNet 的特征进行了可视化，如图 3.9 所示。我们选取带有 2 次循环的 CircleNet-2 进行分析。其中第 1 列是原始输入图像，由于在原始图像中待检测的行人尺度过小，所以我们对图片关键部分进行了放大和裁剪。第 2 列即为放大和裁剪后的图像。第 3 列是第 1 次循环中产生的特征，第 4 列是第 2 次循环中产生的特征。从图 3.9 第 1 行中可以看出，有 2 个行人并排而行，左侧行人清晰可见，而右侧行人被障碍物遮挡。被遮挡的行人在第 1 次循环中没有很好的被激活，然而在第 2 次循环中特

征得到了明显加强。在第 2 个样例中，我们发现不同循环中，网络可以关注不同外表的样本。左侧被遮挡的样本在第 1 次循环中特征较强，而右侧被遮挡的行人在第 2 次循环中特征较强，这可以说明我们提出的环状循环网络具有特征适应能力。第 3 行的图片表示，在第 1 次循环中，学习的特征激活了一些背景区域；而第 2 个次循环中，学习特征更好地激活了行人边界。

图 3.9 在不同循环中特征适配的可视化。第 1 列是原始图片；第 2 列是剪切放大的图片；第 3 列和第 4 列分别是 Circle-1 和 Circle-2 的特征可视化。

Figure 3.9 Visualization of feature adaptation. The first column shows the original image. A cropped patch with a pedestrian is presented in the second column. The visualization of the feature map that from Circle-1 can be seen in the third column and the fourth column shows the Circle-2.

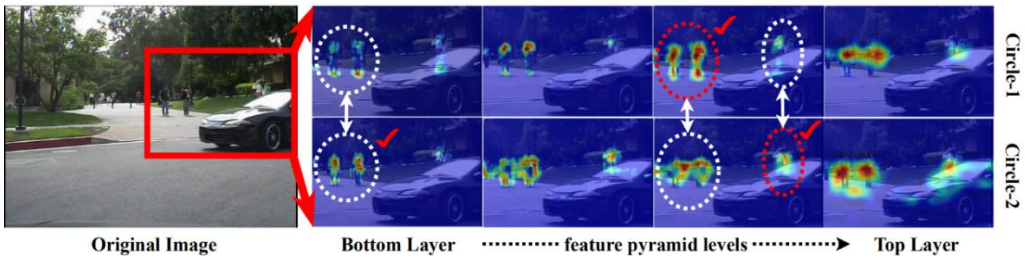


图 3.10 在不同深浅层中特征适配的可视化

Figure 3.10 Visualization of feature adaptation across circles.

图 3.10 进一步展示了 CircleNet-2 特征适配的有效性。图中右侧第 1 行可视化了 CircleNet-2 第 1 次循环产生的特征图，我们叫做 Circle-1；第 2 行可视化了 CircleNet-2 第 2 次循环产生的特征图，叫做 Circle-2。从左到右展示了在同一个循环中，由底层到高层的特征图，即特征金字塔的特征。底层能比较好地激活小

目标，顶层可以较好地激活大目标。对于遮挡的行人，Circle-2 激活了更多的行人细节。如 Circle-2 的第 3 列激活图中，更好地激活了右侧被车辆所遮挡的行人。

在测试阶段，我们从测试集中随机选取了 2792 个行人样本，并且根据遮挡率（None、Partial 和 Heavy）绘制了检测的统计结果，如图 3.11 所示。可以看出 Circle-1 和 Circle-2 在 None 和 Partial 子集上检测的数量相近，但在 Heavy 子集上 Circle-2 的检测数量明显高于 Circle-1。Circle-2 检测到了 512 个严重遮挡的行人，而 Circle-1 只检测到了 128 个。这说明增加循环次数，可以明显提升网络在严重遮挡样本上的检测能力。

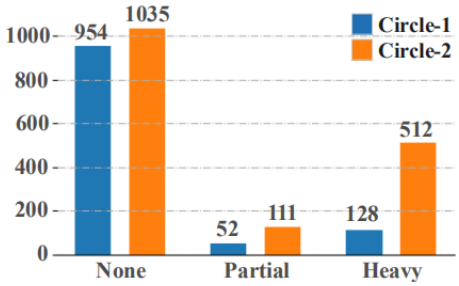
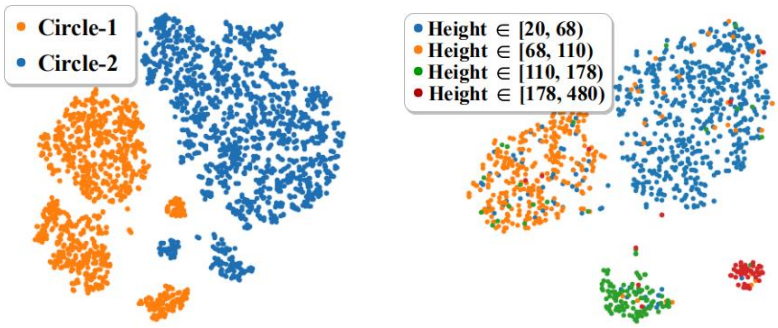


图 3.11 CircleNet 检测结果统计。“None”和“Partial”集合上 Circle-1 和 Circle-2 检测数量相近，但是 Circle-2 可以检测到更多的“Heavy”遮挡对象。

Figure 3.11 Detection result statistics. Circle-1 and Circle-2 detect a comparable number of objects with occlusion ratio None and Partial, but Circle-2 detects significantly more Heavy occlusion objects.



(a) (b)

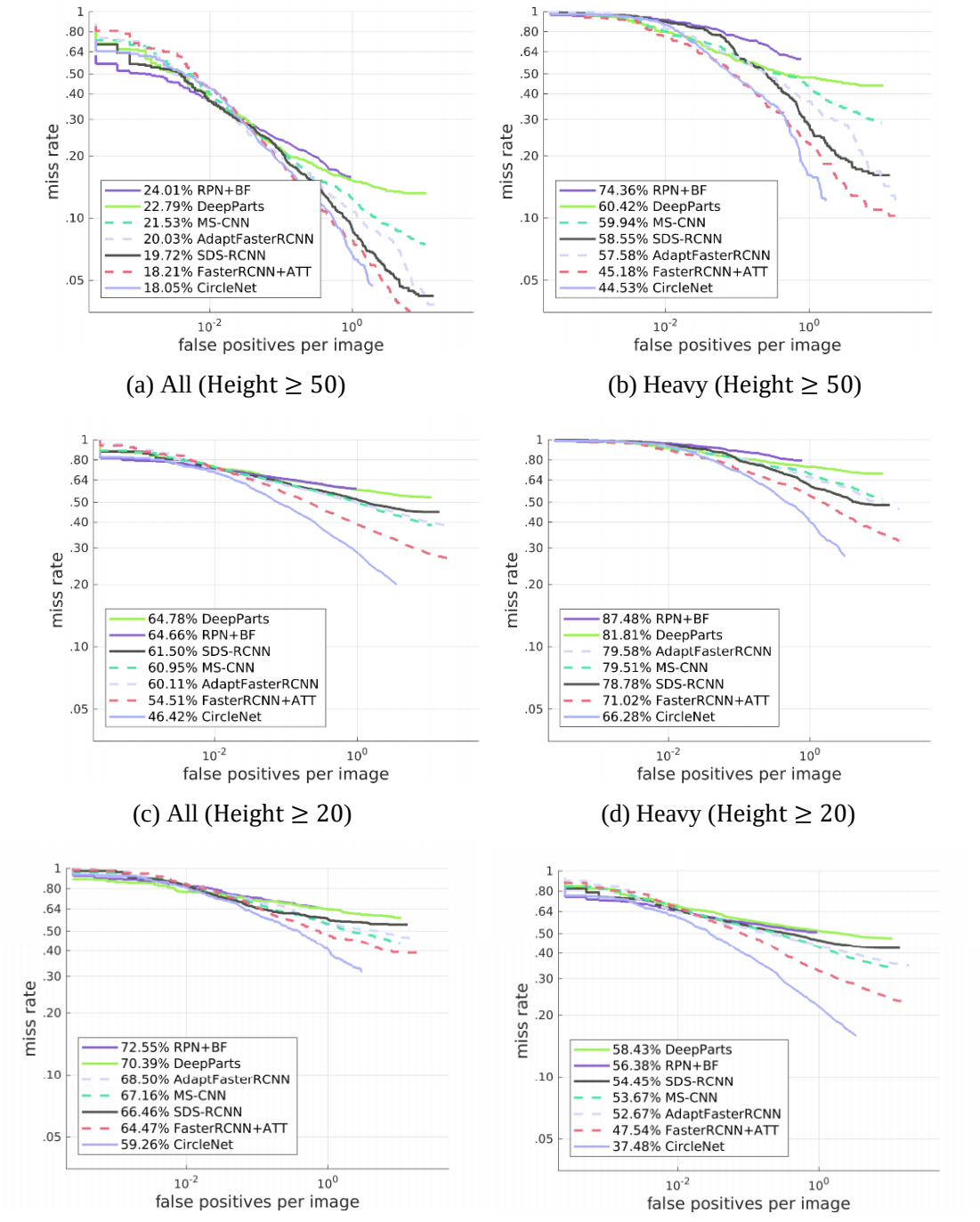
图 3.12 不同特征向量的 t-SNE[77][78]可视化。（a）来自不同循环的行人实例特征，（b）具有不同分辨率的行人实例特征。

Figure 3.12 The t-SNE[77][78] visualization of different feature embedding. (a) Instances from different circles. (b) Instances with different resolution.

我们使用 t-SNE[77][78]可视化了这些行人样本特征嵌入向量，如图 3.12 所示。通过图 3.12（a）可以看出 Circle-1 和 Circle-2 的特征呈现不同的聚类分布，它们

分别对应不同外表的行人实例。3.12（b）表明，不同分辨率的行人有着不同的聚类。而 CircleNet 作为一种分类器集成方法，可以将不同聚类都处理的很好。

3.4.5 与其他当前方法对比



(a) All (Height ≥ 50)　(b) Heavy (Height ≥ 50)

(c) All (Height ≥ 20)　(d) Heavy (Height ≥ 20)

(e) Partial (Height ≥ 20)　(f) None (Height ≥ 20)

图 3.13 在 Caltech 数据集上的性能对比曲线。曲线越低表示性能越好。

Figure 3.13 Performance comparison on the Caltech dataset. Lower curves indicate better performance.

在图 3.13 中，我们与 DeepParts[81]、RPN+BF[44]、SDS-RCNN[80]、MS-CNN[82]、AdaptFasterRCNN[46]和 FasterRCNN+ATT[56]等几种行人检测方法进行了对比。在 Caltech 验证集上， Height ≥ 50和Height ≥ 20两种条件下，丢失率和 FPPI 曲线表明我们的方法均优于其他几种方法。尤其对于小样本行人目标，提升效果明显，如图 3.13（c）到图 3.13（f）在Height ≥ 20集合上的性能曲线。

除了定量分析之外，我们也做了一些定性分析的实验。如图 3.14 所示，相比于其他两个检测器（FasterRCNN + ATT[56]和 MS-CNN[82]），CircleNet 产生了可靠的检测结果，而其他两种方法都出现了漏检行人的情况。这些检测结果在各种遮挡模式下都与标注框非常吻合。并且对于检测到的目标，我们的检测结果在定位上比起另外两种方法更加的精确。

(a) MS-CNN[82]

(b) FasterRCNN+ATT[53]

(c) CircleNet

图 3.14 CircleNet 和其他方法的检测效果对比。该检测结果是在 FPPI=0.1 条件下进行的可视化。绿色框表示标注框；红色框表示检测结果。

Figure 3.14 Qualitative detection results of cropped image patches at FPPI=0.1 on the Caltech test set. The green solid boxes indicate ground truth; the red boxes denote detection results.

表 3.6 在 Caltech 测试集上 CircleNet 和其他方法性能对比

Table 3.6 Comparison of CircleNet with other state-of-the-art methods on the Caltech test set.

Method	Height ≥ 50					Height ≥ 20			
	All	None	Partial	Heavy	Reasonable	All	None	Partial	Heavy
DeepParts	22.79	10.64	19.93	60.42	11.89	64.78	58.43	70.39	81.81
MS-CNN	21.53	8.15	19.24	59.94	9.95	60.95	53.67	67.16	79.51
RPN+BF	24.01	7.68	24.23	74.36	9.58	94.66	56.38	72.55	87.48
AdaptFaster RCNN	20.03	7.01	26.55	57.58	9.18	60.11	52.67	68.50	79.58
SDS-RCNN	19.72	**5.95**	**14.86**	58.55	**7.36**	61.50	54.45	66.46	78.78
FasterRCNN+ATT	18.21	8.46	22.29	45.18	10.33	54.51	47.54	64.47	71.02
CircleNet (Ours)	**18.05**	8.42	20.27	**44.53**	10.21	**46.42**	**37.48**	**59.26**	**66.28**

在表 3.6 中，我们在 Caltech 行人数据集测试集上对比了 state-of-the-art 的方法和 CircleNet 的性能。MS-CNN、RPN+BF、AdaptFaster-RCNN 和 SDS-RCNN 在“Reasonable”集合上实现了很高的性能，但是在遮挡样本和小目标行人样本上性能较低。在低分辨率和遮挡的场景中，CircleNet 明显优于所有对比的方法。在Height ≥ 50 的“All”集合上实现了 MR^{-2}=18.05 的性能，同时保持了在“Reasonable”集合上的性能。对于Height ≥ 50的“Partial”子集，CircleNet 优于 FastRCNN-ATT 方法 2.02%（20.27% vs. 22.29%）。在Height ≥ 20的“All”子集中，CircleNet 性能优于 FasterRCNN-ATT 方法 8.09%（46.42% vs. 54.51%）。CircleNet 在“Partial”和“Heavy”遮挡子集中，也明显好于 FasterRCNN-ATT 方法。我们的方法在在Height ≥ 20所有子集中，都达到了最优的水平。由此表明，相比于其他方法，CircleNet 在遮挡样本和小目标行人样本上可以取得更好的性能。

除 Caltech 之外，我们还在 CityPersons 验证集合上比较了 CircleNet 和其他当前行人检测方法的检测结果，如表 3.7 所示。可以看出，CircleNet 明显优于最新的 OR-CNN 和 FasterRCNN+ATT 在论文[46][53]和[59]中汇报的结果。相比于 Caltech 数据集，CityPersons 数据集包含更接近自动驾驶的真实场景，且图像分辨率更高，更具有挑战性。因此相比于其他两种方法，CircleNet 更具有实际应用价值。

表 3.7 在 CityPersons 验证集上 CircleNet 和其他方法性能对比（Height ≥ 50）。*表示论文[59]中使用的实验设定。

Table 3.7 Comparison of CircleNet with other state-of-the-art methods on the CityPersons validation set. (* denotes the experimental protocols used in [59].)

Method	*Reasonable*	*Heavy**	*Partial**	*Bare**
Adapted Faster RCNN[46]	12.8	-	-	-
Repulsion Loss[59]	11.6	55.3	14.8	7.0
OR-CNN[53]	11.0	51.3	13.7	5.9
CircleNet (Ours)	11.7	**50.2**	**12.2**	7.1

为了直观地评测我们提出方法的检测结果。我们从 Caltech 和 CityPersons 数据集中选取了一些典型的场景，并且可视化了最终的检测结果。在图 3.15 中，我们可视化了 CircleNet 在 Caltech 数据集上的检测结果。我们发现该方法在检测距离摄像头远处的小目标行人上可以检测地非常好，同时对于遮挡严重的样本，检测效果也非常鲁棒。图 3.16 呈现了一些在 CityPersons 数据集中拥挤场景和低分辨率场景的检测结果。在第 1 行第 2 列、第 3 行第 1 列和第 4 行第 1 列的图片中，有许多低分辨率的行人目标，对此我们的检测器几乎不存在漏检的情况。除此之外，在一些其他遮挡场景中，我们的检测器也工作的很好，如图 3.16 的第 4 行第 1 列图像。

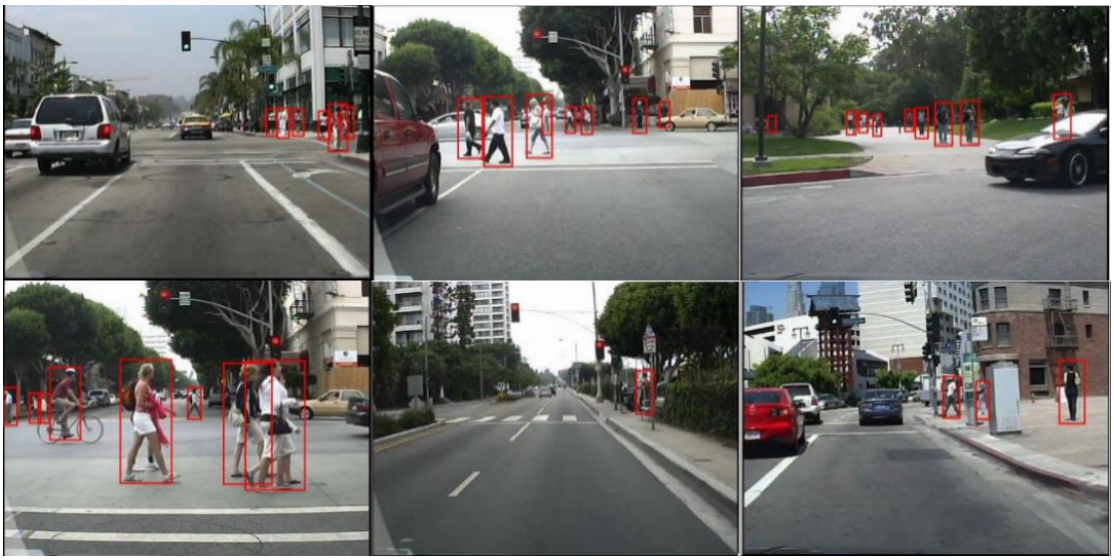


图 3.15 CircleNet 在 Caltech 数据集上的检测结果可视化。红色边界框是在阈值 0.7 条件下预测的行人。

Figure 3.15 Detection examples from the Caltech dataset. Red bounding boxes are predicted pedestrians with threshold 0.7.

图 3.16 CircleNet 在 CityPersons 数据集检测结果可视化。红色边界框是在阈值 0.7 条件下预测的行人。

Figure 3.16 Detection examples from the CityPersons dataset. Red bounding boxes are predicted pedestrians with threshold 0.7.

3.5 本章小结

行人检测是一个非常有挑战的研究方向，尤其是在解决难样本（低分辨率和遮挡目标）的检测中。本章提出来一个新的自适应特征增强模型，我们称之为环状循环网络，用以解决行人检测难样本的问题。环状循环网络模拟了人们尝试识别难样本的时候，往往需要多次进行识别的过程，可以在不同循环上多次检测目标。它通过自上而下和自下而上两个通路多次融合，实现特征自适应。这些环状的循环不仅提升了特征的表达能力，同时也针对不同外表的行人提出有适应性的卷积特征。不同分辨率的行人检测和不同遮挡情况的行人检测都是行人检测中的

难点问题。CircleNet 可以同时实现样本在分辨率和遮挡率上的分解，最后结合行人实例分解训练策略，发挥出了环状循环网络的最大能力。和基准方法 FPN 对比，我们的环状循环网络的性能有明显提高，这表明了我们方法的有效性。在视觉行人检测系统中，我们的方法体现出巨大的潜力，并且为遮挡和低分辨率目标检测提供了更有效的特征表达。

第4章　基于特征校准与增强的行人检测

在本章中，我们首先对遮挡行人检测问题进行了分析，找到解决遮挡行人检测问题的关键因素。然后介绍我们提出的特征校准网络（Feature Calibration Network，FC-Net）整体结构，给出其包含的各个组件。接下来提出一种自激活方法，它通过重用检测网络的分类权重来提供一种简单而有效的方法来估算行人激活图。其次，我们设计了特征校准模块，其旨在突出可见部分特征并抑制行人的被遮挡部分特征。结合特征校准模块可以将深度检测网络升级为特征校准网络，整个网络结构的变化没有引入任何额外的网络参数。最后我们在常用的行人检测基准上应用 FC-Net，并实现了最新的检测性能。除此之外，我们还验证了 FC-Net 在一般目标检测中的适用性。

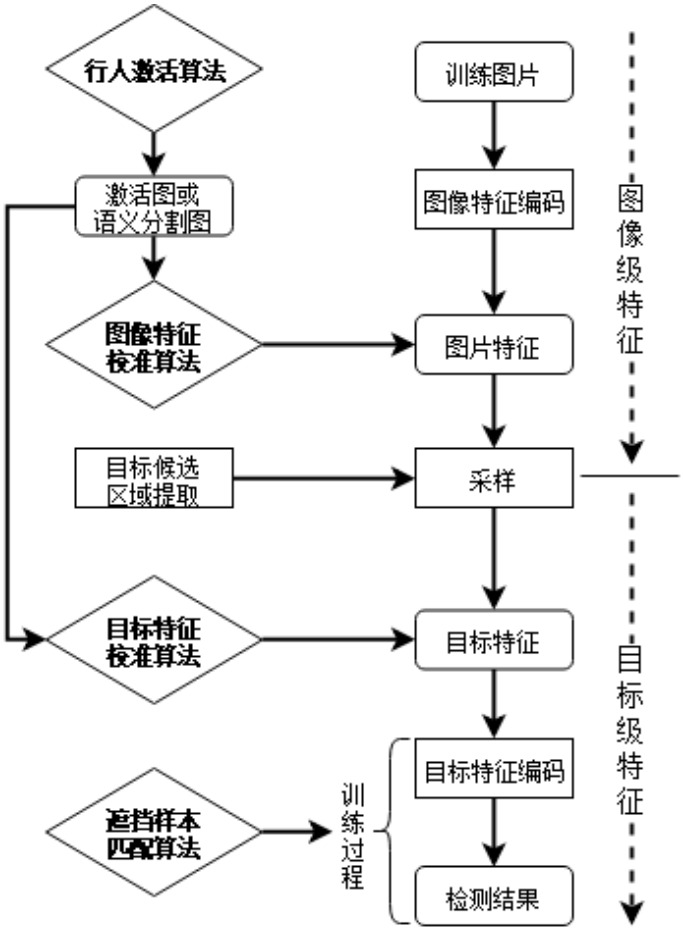


图 4.1 特征校准行人检测流程示意图

Figure 4.1 The framework of feature calibration network

本文提出的行人检测特征校准网络的系统流程如图 4.1 所示，其中主要包括行人激活算法、图像特征校准算法和目标特征校准算法。感兴趣区域采样（RoI Pooling）操作之前，网络中的特征都是图像级的特征，之后是目标级的特征。所提出的自激活（Self-activation，SA）和特征校准（Feature Calibration，FC）模块，可以使卷积特征适应各种遮挡的行人。SA 模块利用了行人局部视觉模式和卷积特征通道之间建立起来的对应关系，这种关系可以反应在在网络训练期间构造的分类器权重向量上。因此，我们使用分类器权重向量与特征图相乘的方式收集跨通道的视觉模式，计算出行人激活图，该过程无需引入任何额外的参数。生成的激活图被进一步送到 FC 模块以增强或减弱基于像素和基于区域的卷积特征。将 SA 和 FC 模块与深度检测网络集成在一起便形成了我们的特征校准网络（FC-Net）。在每次学习迭代中，FC-Net 都会更新分类器权重，这些权重将被重复使用以迭代地校准特征。校准的关键思想是利用行人激活图作为指示器来增强行人可见部分中的特征，同时抑制被遮挡的行人区域中的特征。通过多次特征校准，FC-Net 可以以渐进的方式学习区分行人外表的特征。

4.1 问题分析

解决遮挡行人检测问题需要关注两个关键子问题。第一个子问题是如何增加网络对行人可见区域的关注度，第二个子问题是如何减少来自遮挡区域的噪声干扰。下文将对这两部分分别进行阐述。

对于增加网络对行人可见部分的关注度。Zhou[52]提出了 PDOE+RPN 方法，通过同时回归行人全身位置和行人可见部分位置来完成行人检测和对遮挡程度的估计；Faster-RCNN+ATT[56]方法中使用了可见框注意力网络和部件注意力网络，使其分别学习可见区域框信息和人体关键点信息；MGAN[83]（Mask-Guided Attention Network）是一种掩膜引导的注意力网络，可以利用一个子网预测行人可见部分掩膜。

针对减少噪声干扰。Zhang[53]提出在 OR-CNN 方法中使用遮挡处理单元来屏蔽来自遮挡区域的特征。该遮挡处理单元输出的数值可以作为门限开关，决定最终目标的特征是否融合该遮挡区域。

本文所提出的特征校准网络（FC-Net）可同时解决行人关注度和噪声干扰这两个问题，并且对比上述方法，我们提出的方法不需要额外的标注信息（比如行

人可见部分区域），同时也不需要增加额外的网络参数。它可以天然地融入到各种检测框架中，是一种显式的建模过程，可以学到语义信息非常强的激活图。

4.2 特征校准网络

行人特征校准网络（Feature Calibration Network ，FC-Net），其结构如图 4.2 所示。整体结构中包含：一个主干网络、一个自激活模块（SA）、一个特征校准模块（FC）和一个预测器。FC-Net 以一种新的自激活方式加强行人可见部分特征，并且抑制行人遮挡部分的特征。SA 模块通过重用分类器权重来生成行人激活图，而无需涉及任何其他参数。因此产生了一个极其简约的增强特征语义的方式，同时 FC 模块以基于像素的方式和基于区域的方式校准行人的卷积特征。

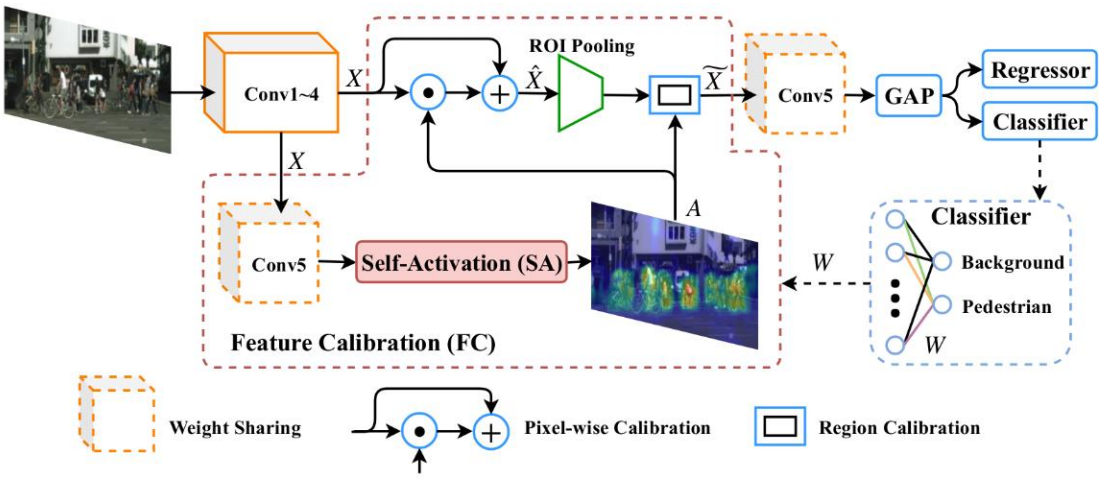


图 4.2 特征校准网络结构示意图。它由一个检测网络、自激活模块和特征校准模块构成。经过多次特征校准，FC-Net 学习到的特征可以增加可见部分，同时抑制被遮挡的部分。在网络中，GAP 表示全局平均池化层。

Figure 4.2 The architecture of the feature calibration network (FC-Net), which is made up of a deep detection network, a self-activation (SA) module, and a feature calibration (FC) module. With multiple iterations of the feature calibration, FC-Net learns features which highlight the visible parts and suppress the occluded regions of pedestrians. In the network, GAP stands for global average pooling.

我们使用了经典的目标检测框架 Faster R-CNN[45]作为我们的基准方法。在它的基础上，我们参考论文 [46]做了一些改进，使得该方法可以有效适用于行人检测问题。首先，我们采用 ResNet50 作为我们的骨干网络，用于图像特征的提取。ResNet50 相比于 VGG16，有更少的参数，并且速度更快。我们还将 Conv4

的步长从 2 更改为 1，这样可以获得更高分辨率的行人特征，有利于小尺寸行人目标的检测。除此之外，我们还重新量化了候选区域提取网络锚点框的参数，让其更好的适用于多尺度行人检测。全局池化层（Global Average Pooling, GAP）用于连接 ResNet 网络的后端，用于将三维的卷积特征转化为向量，方便后端分类器和回归器使用。

4.2.1 行人检测激活图

受到类别激活图[84]（Class Activation Maps，CAM）的启发，我们提出了行人自激活方法（Self-activation，SA）。CAM 之前被用目标定位任务，该方法可以发现该类别中响应最强的位置。我们将该方法扩展到检测网络中，并且使得该激活方法不仅仅局限于类别响应最强的位置，而是激活目标的整体区域。通过该自激活方法产生的激活图，我们称之为行人激活图（Pedestrian Activation Map）。

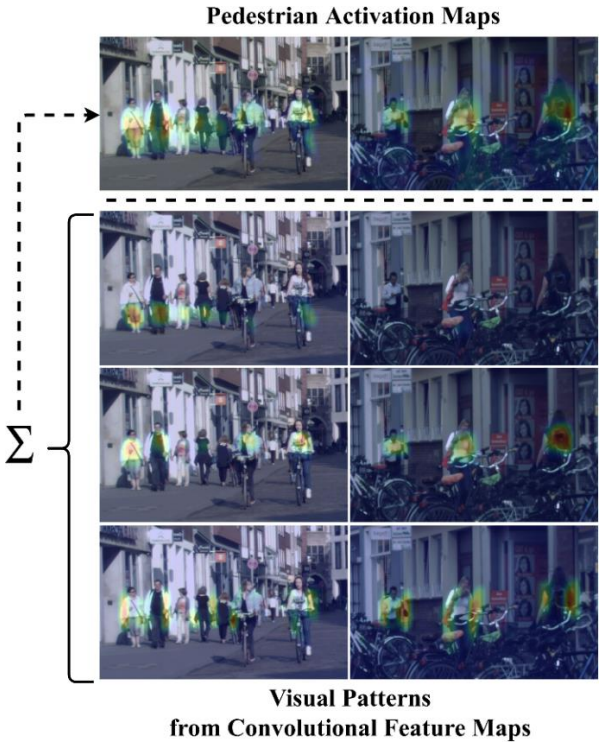


图 4.3 卷积特征的视觉模式。上图为行人激活图，下图为视觉模式。

Figure 4.3 Pedestrian activation maps (upper) and visual patterns (lower).

深度卷积神经网络中实际上包含了学习到的目标的模式，我们称卷积网络的每个通道的输出所激活的区域为一种视觉模式。该视觉模式可能表示目标的一个关键点、身体部位或者一种组合方式。通过融合这些视觉模式，可以得到行人激活图。该激活图可以识别行人区域，具有加强行人可见部分特征和抑制遮挡区域

特征的特性。如图 4.3 中后 3 行，我们可视化了一些行人视觉模式，其中左侧为非遮挡行人，右侧为遮挡行人。通过观察我们可以发现，在图第 2 行中，该视觉模式可识别行人腿部区域，而在右侧行人腿部被遮挡的情况下，该视觉模式没有被激活。

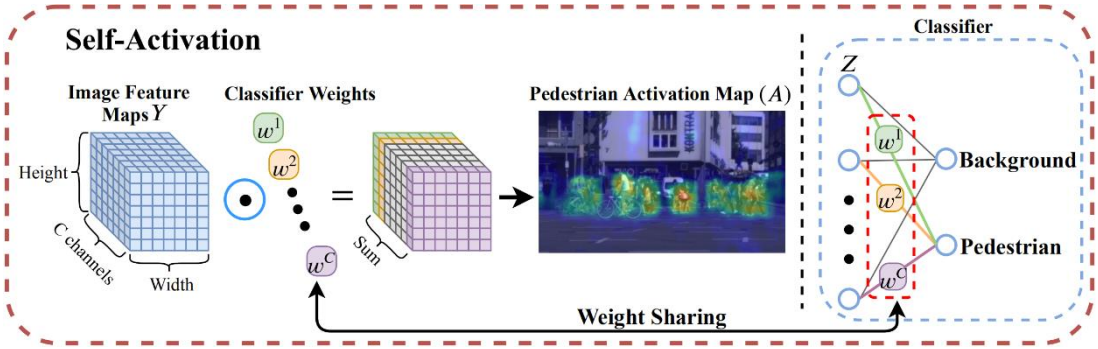


图 4.4 自激活模块。通过重用分类器权重向量，它将特征图从多个通道转换为行人激活图。

Figure 4.4 The self-activation (SA) module. It converts feature maps from multiple channels to a pedestrian activation map by reusing the classifier weight vector.

在检测网络中的最后一个卷积层上，每个候选区域的卷积特征在通过全局池化层（Global Average Pooling）之后，将输出一个 C 维的向量，这里 C 表示特征通道的数量。这样一个向量通过全连接层和 Softmax 操作，可转换成一个类别置信度。

自激活模块是网络的一个分支，如图 4.4 所示。它复用了分类器的权重，可对所有通道的卷积特征进行加权求和后得到行人激活图。行人检测问题是二分类问题，所以它的分类器的全连接层中包含两组参数，一组对应行人类别，另外一组对应背景类别。我们定义对应行人类别的权重为 $W = (w^1, w^2, \dots, w^C)^T \in R^C$，它被用来生成行人激活图。我们定义$Y \in R^{M \times N \times C}$表示输入图像的卷积特征，这里 M 和 N 分别表示特征的宽度和高度。在行人激活图上的一个元素$A_{m,n}$的计算方式如下：

$$A_{m,n} = \sum_{c=1}^{C} w^c \cdot Y_{m,n}^c \tag{4.1}$$

其中 m 和 n 表示特征图上的 2D 像素坐标，c 表示特征通道的序号，并且$A \in R^{M \times N}$。

对于一个行人目标，因为卷积滤波器是针对不同的视觉模式而学习的，所以不同的特征通道对行人的不同部分敏感。受益于 RoI Pooling 不改变特征通道的顺序的事实，检测器的学习过程构建了特征通道和分类器权重向量之间的统计关

系。权重越大，相应的特征通道信息越丰富；反之，信息则越匮乏。通过使用公式 4.1，当相应的特征通道和权重具有较大的值时，我们可以将视觉模式聚合成行人激活图。当相应特征通道或权重的值较小时，我们还可以用来抑制其遮挡区域。

4.2.2 像素级特征校准

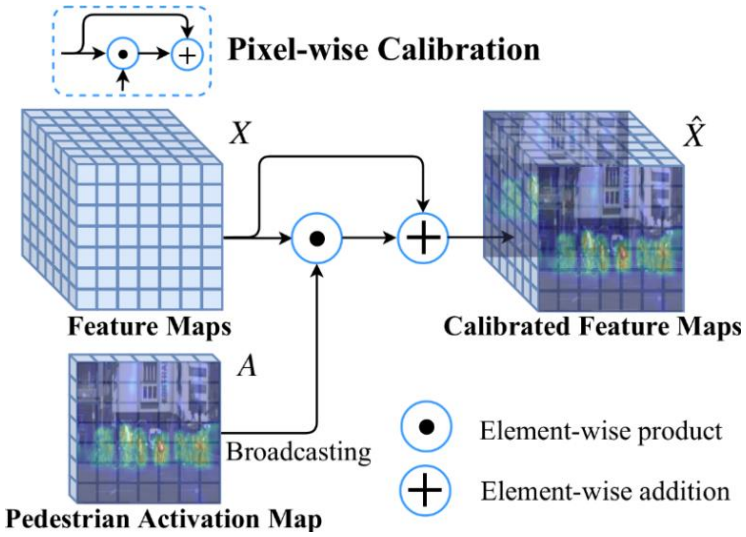


图 4.5 像素特征校准。它利用行人激活图来校准特征图。它的目标是在遮挡区域上抑制具有高值的特征通道，同时加强在可见部分上的特征值。

Figure 4.5 Pixel-wise feature calibration. It leverages the pedestrian activation map to calibrate the feature maps. It targets suppressing the feature channels of high values on occluded regions while aggregating those of high values on visible parts.

为了利用行人激活图中所包含的信息，我们采用了特征校准过程以有效地处理遮挡的情况，进而优化卷积特征。所选择的特征校准应具备两种能力。首先，它应该有能力选择空间中的遮挡位置。特别是，它必须能够抑制在遮挡区域上输出高特征值的通道和位置。其次，它应该包含一些上下文信息，以便当行人的重要部分被遮挡时，区域级的目标特征也可用于检测行人。为了满足这些要求，我们设计了像素级特征校准和目标级区域特征校准。前者强制卷积特征关注行人的可见部分和有判别性的部分，而后者利用行人激活图通过引入多级上下文信息来选择最具辨别力的区域。

像素级特征校准具体执行过程如图 4.5 所示，它是通过一个像素级的元素点乘操作和一个相加操作来实现的。我们定义校准前和校准后第 c 通道的特征图分

别为$X = \{X^c\}$和 $\hat{X} = \{\hat{X}^c\}$。像素级特征校准操作被定义为：

$$\hat{X}^c = A \odot X^c + X^c, c = 1, \ldots, C \tag{4.2}$$

这里⊙表示元素点乘。这个校准后的特征将被送入后端，用于行人候选区域特征的提取。元素点乘操作将行人激活图反映的遮挡和非遮挡置信度传递给每个特征通道。但是行人激活图也不一定准确，因为行人有着各种各样的外表和杂乱的背景。加法运算的使用保持了原始特征，从而平滑了像素级特征校准的过程。

4.2.3 目标级特征校准

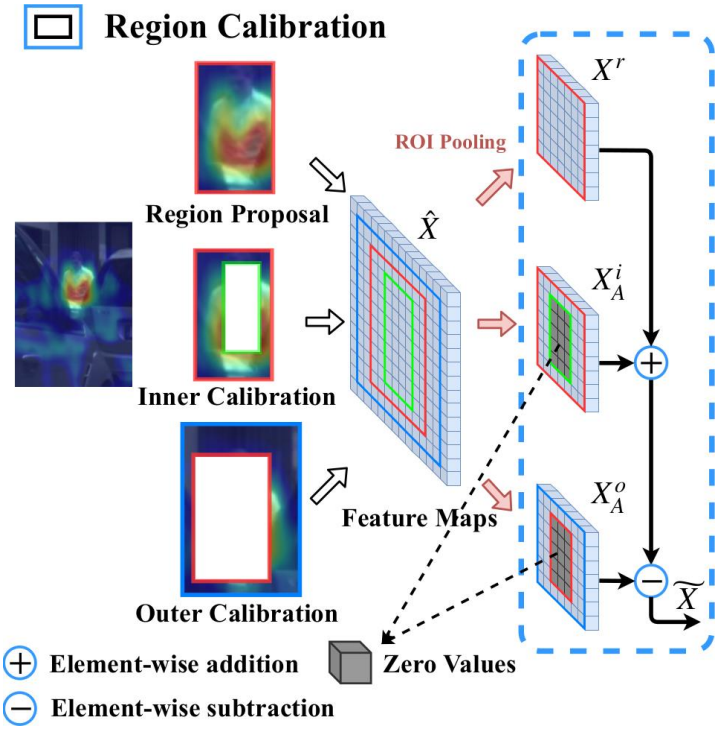


图 4.6 区域特征校准。它利用行人激活图来指导提取行人的上下文特征。

Figure 4.6 The region-based calibration. It leverages the pedestrian activation map to guide extracting contextual features of a pedestrian.

为使目标特征具有极强的行人判别性。在使用行人激活图时，我们提出一种适应性的上下文特征处理模块来加强特征表示，称之为目标级区域特征校准。一些遮挡的行人区域与一些混有复杂背景区域相比，它们的内边缘和外边缘的上下文信息明显不同。如图 4.6 所示，我们给每一个行人候选区域（Region Proposal）分别定义了一个内校准框（Inner Calibration）和一个外校准框（Outer Calibration）。外校准框是一个与行人候选区域相似矩形环，它的外在矩形边界的高和宽定义为 $\text{height} = h \times r$和$\text{width} = w \times r$，这里 h 和 w 是候选区域的高和宽，$r > 1$是一个

超参数，它的内矩形边界是候选区域的高和宽。以相似的方式，我们定义内校准框,它的内矩形的边界定义为$\text{height} = h/r$和$\text{width} = w/r$，外矩形边界是候选区域的高和宽。结合行人激活图的引导，外校准框和内校准框的中心位置可以发生偏移，我们选取候选区域内数值最大的激活图点作为外校准框和内校准框的中心点。

两个校准矩形区域的位置是由行人激活图A所决定的，因此区域校准过程中可以利用行人激活图引导提取具有行人上下文信息的特征。我们用X^r、X_A^i和X_A^o表示在 RoI Pooling 操作之后相同尺度的三种特征。如图 4.6 所示，X^r和X_A^i都来自目标候选区域，但是在X_A^i的内边界以内的特征被置 0，即图中灰色区域；X_A^o来自外校准框，但是在X_A^o以内的候选区域范围内的特征被置 0，即图中灰色区域，这表示该区域不参与特征表达。最终我们定义校准后的特征$\tilde{X}$为：

$$\tilde{X} = X^r + X_A^i - X_A^o \tag{4.3}$$

我们希望X^r和X_A^i包含行人目标特征，而X_A^o不包含行人目标特征。由于X_A^i和X_A^o目标相反，所以这里我们用相减的操作。

4.3 实验结果与分析

在这一节，我们首先可视化并且分析了行人激活图，以及各个通道对应的视觉模式。然后分别分析了像素级特征校准和目标级区域特征校准的有效性，同时给出了算法的运行效率。其次在行人检测研究中广泛应用的 CityPersons 数据集和有高质量标注的 Caltech 数据集[85]上进行了评测和对比。最后我们将方法推广至通用目标检测方向，并且在 PASCAL VOC 2007 数据集上进行了评测。验证了我们模型的有效性，同时分析了在少数类别上性能下降的原因。

4.3.1 实验设定

我们使用 ResNet50 作为主干网的 Faster RCNN 为基准。在 Caltech 数据上评测时，我们使用了修正的高质量标注[85]，同时我们将图像分辨率上采样到900×1200，并且使用 CityPersons 数据集进行预训练。我们使用 8 个 Nvidia V100 GPU 训练网络。对于 CityPersons 数据集，模型总共训练了 6k 次迭代。初始学习率设定为 0.008，在 5k 次训练迭代后，学习率衰减 10 倍，优化器使用的随机梯度下降算法（Stochastic Gradient Descent，SGD）。每个 GPU 上的小批量包含 1 幅图像。权重衰减设定为 0.0001，动量设定为 0.9。交通场景下的行人尺度较小，

论文[46]中提到通过上采样图片可以获得更高分辨率的特征，有益于行人检测。因此，为了和其他行人检测方法进行公平的比较，我们在 CityPersons 数据集上也分别使用了 1 倍和 1.3 倍图像分辨率作为评测标准，并且同时使用 ImageNet 数据集进行预训练。Caltech 数据集和 CityPersons 数据集详细介绍见 2.4 节，评测标准使用对数平均丢失率（MR^{-2}），详细介绍见 2.5 节。

4.3.2 行人激活图可视化

在解决行人检测遮挡问题上，我们提出行人激活图方法。我们首先可视化了几个行人激活的实例，然后展示了两种典型的行人视觉模式。通过激活图的可视化发现该方法可以用于加强行人可见部分特征和抑制遮挡区域特征。在后续试验中，我们还验证了该方法在去除背景干扰上具有一定的作用。

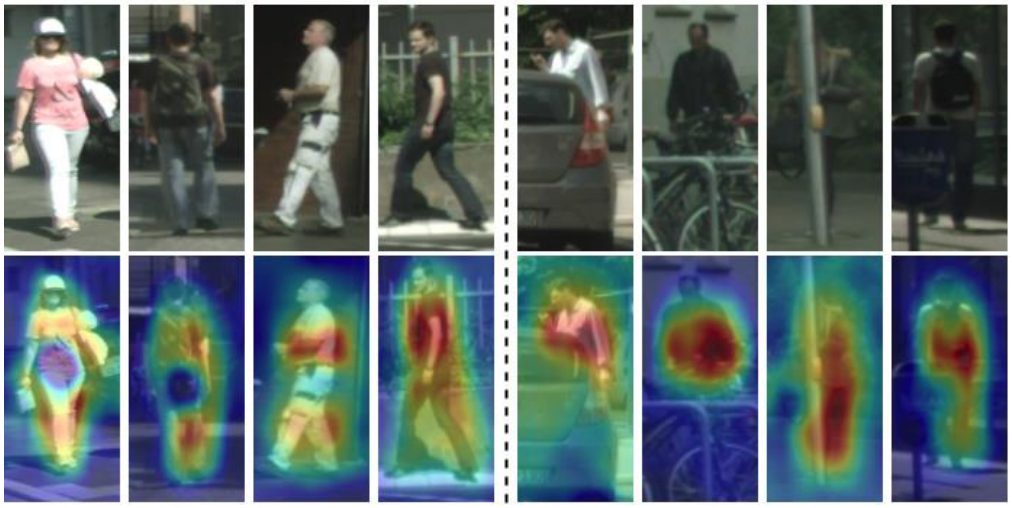


图 4.7 行人激活图可视化。左侧：非遮挡行人实例；右侧：遮挡行人实例。

Figure 4.7 Examples of pedestrians and pedestrian activation maps. Left: non-occluded instances. Right: occluded instances.

图 4.7 展示了一些行人激活图的实例，其中左侧为非遮挡行人，右侧为遮挡行人。可以看出，我们的方法可以自适应地抑制各种遮挡区域，并且加强可见的行人部分。例如图中右侧遮挡行人的第二列，一个行人被自行车等金属物所遮挡，激活图只激活了上半身区域，下半部分呈现抑制状态。再如右侧的第四列，一个背包行人被邮筒所遮挡，被遮挡的区域没有激活，这样对该区域的特征可以有效地进行抑制。

（a）激活模式 1（脚部）

（b）激活模式 2（两侧）

图 4.8 行人激活模式可视化。（a）激活模式 1（脚部），（b）激活模式 2（两侧）

Figure 4.8 Pedestrian visual patterns.

在可视化了整体行人激活图之后，我们想进一步分析每个通道上的行人激活模式。图 4.8 可视化了两个通道的行人激活模式，可以看出图 4.8（a）中的激活模式激活了行人的脚步区域，而图 4.8（b）中的激活模式激活了行人两侧的手臂信息和局部上下文信息。可见这些可视化的激活模式在行人检测中起到了至关重要的作用。

4.3.3 像素级特征校准和目标级特征校准验证

在表 4.1 中，我们对像素级特征校准（Pixel-wise calibration）和目标级区域特征校准（Region calibration）进行了验证。相比于基准方法，在 1.3 倍输入图像尺度下，像素级特征校准方法 Reasonable 子集中较少了 0.78% MR^{-2}，在 Heavy 子集中减少了 3.15% MR^{-2}，在 Reasonable+Heavy 子集上减少了 1.65% MR^{-2}。目标级区域特征校准方法在 Reasonable、Heavy 和 Reasonable+Heavy 子集中分别减少 1.50% MR^{-2}，5.39% MR^{-2} 和 2.75% MR^{-2}。

表 4.1 在 CityPersons 验证集上特征校准模块的消融实验性能比较。更小的数值表示更好的性能。

Table 4.1 Ablation study of the proposed feature calibration (FC) module on the CityPersons validation dataset with MR^{-2}. Smaller number indicates better performance.

Method			Scale	*Reasonable*		*Heavy*		*Reasonable+Heavy*	
	Pixel-wise calibration	Region calibration		MR	ΔMR	MR	ΔMR	MR	ΔMR
FC-Net			× 1	15.18	-	51.05	-	31.05	-
	√	√	× 1	13.93	+1.25	46.79	+4.26	29.64	+1.41
			× 1.3	13.21	-	46.75	-	29.45	-
	√		× 1.3	12.43	+0.78	43.60	+3.15	27.80	+1.65
		√	× 1.3	11.71	+1.50	41.36	+5.39	26.70	+2.75
	√	√	× 1.3	11.63	+1.58	42.77	+3.98	26.21	+3.24

除了验证特征校准模块的有效性，我们还测试了添加两个模块之后的模型运行速度。在表 4.2 中，将 FC-Net 的测试效率与 Faster R-CNN 基准进行了比较，在测试阶段，一次调用 SA 和 FC 模块，时间由原来的 0.248s 增加到了 0.290s，每个图像仅增加 0.042 秒的运行时间。可见增加的 SA 和 FC 模块只牺牲了 FC-Net 很小的运行效率，检测性能却得到了有效提升。

表 4.2 在 CityPersons 数据集上测试速度比较

Table 4.2 Comparison of detection speed on CityPersons.

Method	Inference Time (s/image)
Faster R-CNN	0.248
FC-Net with SA	0.283
FC-Net with FC	0.284
FC-Net with SA+FC	0.290

4.3.4 校准模块中超参数的选择

内部和外部校准矩形的中心位置是通过在激活图上搜索最大值的区域来确定的。然而内部和外部校准矩形的缩放比率参数（r）需要根据经验来确定。如表 4.3 所示，通过在$[1.0, 2.0]$范围内搜索，我们可以观察到在高度缩放比率$r = 1.8$时可获得最佳性能。由于宽度的比例可能与高度的比例不同，所以在使用最佳高

度比例的前提下，我们进一步确定最佳的宽度比率为 1.0，如表 4.4 所示。最佳宽度比率小于最佳高度比率，其原因可能是在行人候选区域的周围可能存在水平方向上的其他行人，这可能会影响检测器的判断。但是，像素级校准不依赖任何上下文信息，因此即使在拥挤的场景中也有效。

宽度缩放比和高度缩放比是区域校准模块中的两个超参数。表 4.3 和表 4.4 中的消融实验表明，垂直方向的上下文信息比水平方向的上下文信息更重要。原因可能是在垂直方向上行人和背景之间存在更多的共存（共生）信息，比如行人和街道，或者行人和车辆，但是行人和天空就不能存在共生信息。当水平方向上有多个行人时，上下文信息可能会受到干扰。

表 4.3 不同垂直方向缩放比率下的 MR^{-2} 性能比较。此缩放比为外校准框和候选区域的高度比例，同时也是候选区域和内校准框的高度比例，并且此时固定宽度缩放比为 1。

Table 3.3 With the width ratio = 1, MR^{-2} under different ratios for height between the outer rectangle and the region proposal, and between the region proposal and the inner rectangle (see Fig.4.6).

Ratio (Height)	1.0	1.4	1.6	1.8	2.0
Reasoanble	13.21	13.00	12.24	**11.71**	12.78
Heavy	46.75	43.34	42.10	**41.36**	42.09
Reasonable+Heavy	29.45	27.75	27.13	**26.70**	27.11

表 4.4 不同水平方向缩放比率下的 MR^{-2} 性能比较。此缩放比为外校准框和候选区域的宽度比例，同时也是候选区域和内校准框的宽度比例，并且此时固定高度缩放比为 1.8。

Table 3.3 With the height ratio = 1.8, MR^{-2} under different ratios for width between the outer rectangle and the region proposal, and between the region proposal and the inner rectangle (see Fig.4.6).

Ratio (Width)	1.0	1.4	1.6	1.8
Reasonable	**11.71**	12.09	12.90	12.87
Heavy	**41.36**	42.51	43.29	42.98
Reasonable+Heavy	**26.70**	27.53	28.23	28.81

4.3.5 方法有效性分析

在行人检测方向的研究中，复杂的背景也一直是影响检测性能的一个主要因素，为了验证像素特征校准可以有效抑制背景干扰，我们对错误的检测结果进行

了分析。首先，我们对“背景类错误”进行定义，即当一个检测结果和任意的标注框的交并比（IoU）小于 0.2 的时候，我们就称这个错误的检测结果属于背景类错误。可以使用背景类错误占所有错误检测结果的比例值来验证我们方法的有效性，如下图所示。在图 4.9 中，蓝色曲线表示基准方法的背景类错误占整体错误的比例，由图可见背景错误占比非常多，在 FPPI=0.056 到 FPPI=0.316 之间时，背景类错误占到了总错误检测结果的 70%以上，这里 FPPI 表示单位图像下的虚警率。在使用像素特征校准模块之后，如图中黑线所示，背景类错误显著减少，在 FPPI=1.0 时，背景类错误从 58%降到 48%，并且下降趋势在 FPPI=0.316 到 FPPI=1.0 之间尤为明显。这表明在使用像素特征校准之后，背景干扰得到了很好的抑制。

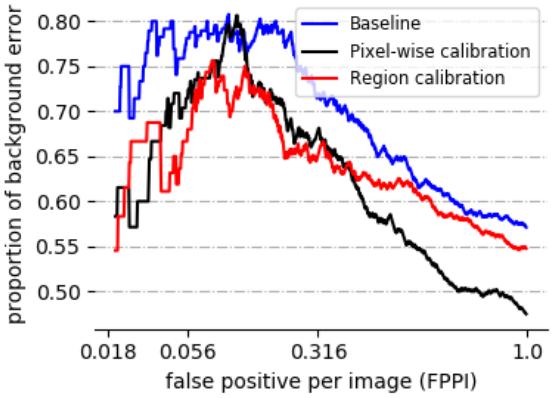


图 4.9 背景类错误比例分析。通过使用特征校准模块，由于背景影响引起的伪正例比例显著减少。

Figure 4.9 Analysis of background error. By applying the proposed feature calibration (FC) module, the proportion of false positives caused by background is significantly reduced.

表 4.5 在 CityPersons 验证集上和 FasterRCNN+ATT 方法对比。FasterRCNN+ATT 是一种注意力引导机制的遮挡的行人检测方法。

Table 4.5 Comparison with the state-of-the-art FasterRCNN+ATT on the CityPersons validation set with MR^{-2}, which is an attention guided approach specified for occluded pedestrian detection.

Method	*Reasonable*	*Heavy*	*Reasonable+Heavy*
FasterRCNN+ATT[56]	15.96	56.66	38.23
FC-Net+ATT	14.82	49.02	31.01
FC-Net (Ours)	13.93	46.79	29.64

为了验证我们方法相比其他注意力机制模型的优势，我们对比了FasterRCNN+ATT[56]方法。该方法是一种解决遮挡行人问题的检测模型，它结合了 SENet[86]中注意力机制，使用一个子网络为每一个卷积特征通道生成一个权重，再用这个权重加权每个特征通道，最后用加权之后的特征进行目标特征提取。而我们的方法使用的是一种显式的方法进行特征通道的加权，每个权重具有可解释性。通过表 4.5 我们可以看出，我们的方法在 Heavy 和 Reasonable+Heavy 子集上，分别优于原始 FasterRCNN+ATT 方法 9.87%和 8.59%。同时在 Reasoanble 子集上，它也有较好的性能。我们在 FC-Net 框架的基础上实现了 FasterRCNN+ATT 中用到的注意力机制模块，我们叫做 FC-Net+ATT。通过对比可以看出，FC-Net 的检测性能要好于我们复现的 FC-Net+ATT，这更说明了我们提出行人激活方法优于 FasterRCNN+ATT 中的注意力机制。

表 4.6 FC-Net 和其他上下文模型性能对比。为了公平的比较，所有的方法都使用了单尺度特征。

Table 4.6 Comparison of context modules. For a fair comparison, all the methods use single scale features.

Method	*Reasonable*	*Heavy*	*Reasonable+Heavy*
MS-CNN[82]	13.22	45.27	28.62
MultiPath[87]	12.19	44.04	27.42
FC-Net (Ours)	11.63	42.77	26.21

区域特征校准模块也可以看作是一种自适应的上下文特征融合模块。为了验证该模块的优势，我们对比了一些其他上下文模型的方法，其中有 MS-CNN[82]和 MultiPath[87]。在表 4.6 中可以看出，我们提出的模块性能优于其他两种方法。因为我们的区域校准在行人激活图的指导下，可以根据行人激活图自适应地产生内部和外部区域。相反，MS-CNN 和 MultiPath 中的使用了预定义的区域，因此无法产生自适应的结果。

4.3.6 和当前行人检测方法对比

我们分别在 CityPersons 验证集和测试集上进行了评测，并且和当前的一些解决遮挡问题的行人检测方法进行了比较，其中有 Adapted FasterRCNN[46]、Repulsion Loss[59]、OR-CNN[53]和 AEMS-RPN[88]。表 4.7 展示了在 CityPersons 验证集上的性能对比。当使用 1.3 倍输入图像尺度的条件下进行测试时，我们的方

法分别在 Heavy 和 Partial 子集上超出 OR-CNN 方法 8.5% MR^{-2} 和 1.8% MR^{-2}。同时在 Reasonable 子集上也保持了一个可以接受的性能。在使用 1 倍尺度测试图像时，它在 Heavy 和 Partial 子集上，比 OR-CNN 减少了 11.4%和 1.8%的平均对数丢失率（MR^{-2}）。

表 4.7 FC-Net 在 CityPersons 验证集上的性能

Table 4.7 Comparison with the state-of-the-art methods on the CityPersons validation set with MR^{-2}.

Method	Scale	*Reasonable*	*Heavy*	*Partial*
Adapted FasterRCNN[46]	× 1	15.4	-	-
	× 1.3	12.8	-	-
Repulsion Loss[59]	× 1	13.2	56.9	16.8
	× 1.3	11.6	55.3	14.8
OR-CNN[53]	× 1	12.8	55.7	15.3
	× 1.3	**11.0**	51.3	13.7
AEMS-RPN[88]	× 1	13.7	-	-
	× 1.3	12.2	-	-
FC-Net (Ours)	× 1	13.5	44.3	14.0
	× 1.3	11.3	**42.8**	**11.9**

表 4.8 FC-Net 在 CityPersons 测试集上的性能对比。我们方法的结果由 CityPersons 的作者进行的评估，比较的其它结果来自 CityPersons 的官方网站[89]。

Table 4.8 Comparison with the state-of-the-art methods on the CityPersons test dataset with MR^{-2}. The results of our approach are evaluated by the authors of CityPersons and the compared results are from the official website of CityPersons.

Method	Scale	*All*	*Reasonable*	*Reasonable_small*	*Heavy*
Adapted FasterRCNN[46]	× 1.3	43.86	12.97	37.24	50.47
Repulsion Loss[59]	× 1.5	39.17	11.48	15.67	52.59
OR-CNN[53]	× 1.3	40.19	11.32	14.19	51.43
FC-Net (Ours)	× 1.3	39.26	12.24	16.67	41.14

为了公平地和公开方法进行对比，我们将结果发送至官网，然后将收到的反馈结果贴入表 4.8。在 CityPersons 未公开标注的测试集上，我们对方法进行评测，该结果是由 CityPersons 的作者进行的官方评估，比较方法的结果来自 CityPersons 的官方网站[89]。从表中我们可以看出，在 Heavy 子集上，FC-Net 的性能优于 OR-CNN 方法 10.29%（41.14% vs. 51.43%）。在 All 这个比较大的行人子集上，它产生了一个不错的性能 39.26%的 MR^{-2}，仅次于使用 1.5 倍测试图像的 Replusion Loss 方法产生的 39.17%MR^{-2}，而我们使用的是 1.3 倍测试图片。由于硬件条件

的限制，我们暂时没有在 1.5 倍测试图像上进行评测。如果在 1.5 倍测试图像上进行测试，将会取得更好的性能，并且有机会超过 Replusion Loss 方法。

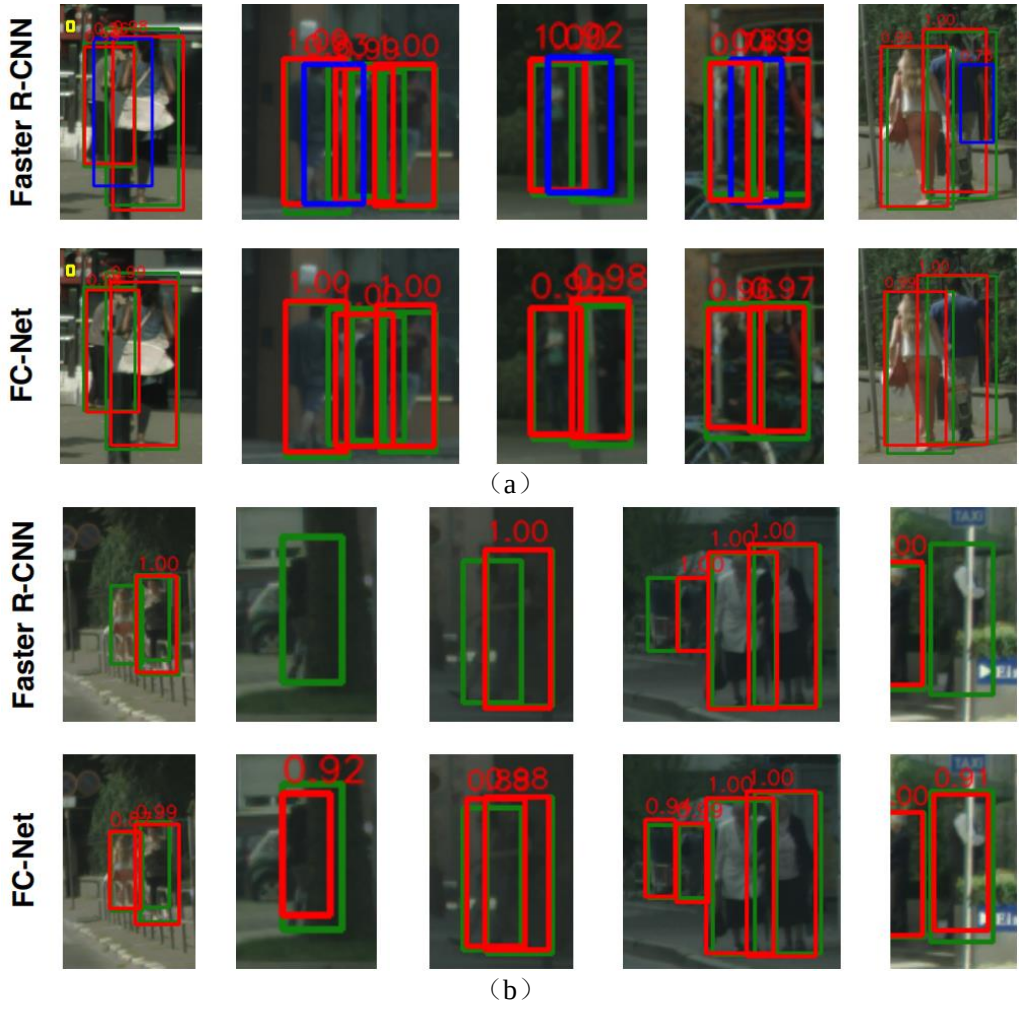


图 4.10 Faster R-CNN 和 FC-Net 在遮挡行人上的检测结果的对比。红色边框表示正确检测到的行人。蓝色框表示误报，绿色框表示标注框。（a）FC-Net 产生较少的误报。（b）FC-Net 比 Faster R-CNN 检测到更多的被遮挡的行人。

Figure 4.10 Comparison of Faster R-CNN and FC-Net on occluded pedestrians. The red bounding boxes indicate correctly detected pedestrians. The blue boxes indicate false positives and the green boxes the ground-truth. (a) FC-Net produces fewer false positives. (b) FC-Net detects more occluded pedestrians than Faster R-CNN.

为了显示我们提出的 SA 和 FC 模块对遮挡处理的有效性，我们展示了更多检测结果。在图 4.10 中，我们比较了 Faster R-CNN 和 FC-Net 在 CityPersons 验证集上的检测结果，该数据集中存在大量的遮挡情况。通过图 4.10（a）可以看出，与 Faster R-CNN 相比，FC-Net 产生更少的误检。这一部分原因是由于像素级特征校准去除了来自杂乱背景的干扰，另一部分原因是区域特征校准加强了行

人的特征表示，使得检测结果更加精确。在图 4.10（b）中，FC-Net 相比于 Faster R-CNN，FC-Net 检测到了更多的遮挡行人样本。这主要是由于行人激活图很好地加强了行人可见部分特征，并且成功抑制了来自遮挡区域的噪声干扰。

图 4.11 FC-Net 在 CityPersons 验证集上的检测结果。红色、蓝色和绿色边界框分别表示检测结果、伪正例和标注框。

Figure 4.11 Examples on the CityPersons dataset. The red, blue and green boxes indicate correctly detected pedestrians, false positives, and ground-truth, respectively.

我们在 CityPersons 数据集中选取了一些典型的拥挤和遮挡场景，对最终的检测结果进行了可视化，如图 4.11 所示。其中红色边界框表示检测结果，蓝色边界框表示误检（false negative），绿色边界框表示行人标注框（ground-truth）。由图可见，FC-Net 可以有效精准地检测遮挡的行人，即使在拥挤场景下该方法的检测结果也很精确。在检测结果中，只存在一些少量的低分辨率行人漏检的情况。我们放大了这些检测结果图片，发现有一些误检是由于数据集标注缺失而导致的。如果能够提供更加干净的标注文件，那可视化中的误检将进一步减少。这样的检测结果已经可以满足一些工业应用场景中的需求，比如行人再识别系统、交通监管辅助系统和智能视频监控场景等。

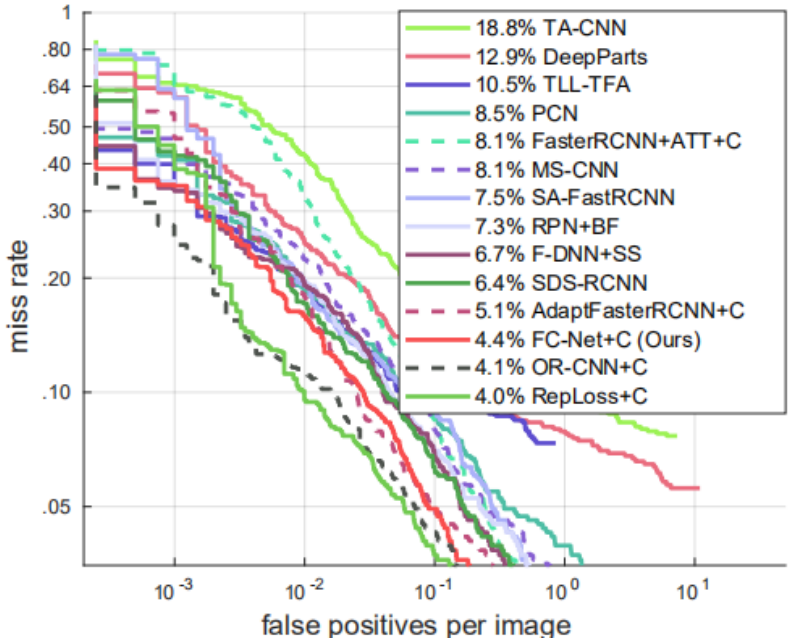


图 4.12 Caltech 数据集测试集上性能对比。“C”表示使用 CityPersons 数据集进行模型的预训练。

Figure 4.12 Comparison with state-of-the-art approaches on the Caltech dataset. “C” indicates models pre-trained on CityPersons. FC-Net achieves 4.4% MR^{-2} and stays on the performance leading board.

在 Caltech 数据集上进行测试。由于原始的 Caltech 数据集的标注存在一些噪声，因此我们使用了修正过标注的高质量的 Caltech 数据集[85]，并同样使用 MR^{-2} 作为评测指标。我们首先在 CityPersons 数据集上进行了模型的预训练，然后在 Caltech 训练集上进行了微调。图 4.12 对比了 FC-Net 与当前主流方法在 Caltech 的 Reasonable 子集上评测结果。如图所示，FC-Net 实现了 4.4%的 MR^{-2}，位于榜单的领先水平，仅仅与 RepLoss+C 相差 0.4%的 MR^{-2}。

4.3.7 在通用目标检测数据集中评测

除了应用于行人检测之外，FC-Net 也可用于通用目标检测。为了验证这一点，我们在 PASCAL VOC 2007[74]数据集上进行了测试。这个数据集包含 20 类通用目标，其中包含日常生活中几个常见的类别，比如汽车、公交车、人、猫、电视和桌子等。我们仍然选用使用 ResNet-50 的 Faster R-CNN 作为我们的基准方法，同时在 ImageNet 数据集上进行预训练。然后在 PASCAL VOC 数据集训练集和验证集上进行微调，最后在它的测试集上面进行评测。从表 4.9 中可以看出，相比于 Faster R-CNN 我们的方法提升了 1.3%的 mAP（73.3% vs. 74.6%），并且

在很多类别的检测上都有提升。在“aero”、“boats”、“sofa”和“train”几个类别上，提升的更为明显，分别提升了 6.3%、5.9%、5.9%和 4.3%。这对于非常有挑战性的通用目标检测任务来说，是很显著的提升。在几个少数的类别上性能稍有下降，例如“bike”，通过可视化其激活图进行分析得知，“bike”的激活图具有许多“孔”，如图 4.13 所示，这可能会引起对目标附近背景区域上的特征像素进行加强，降低了检测性能。

表 4.9 FC-Net 在 PASCAL VOC 2007 上通用目标性能评估

Table 4.9 General object detection results evaluated on PASCAL VOC 2007.

Method	aero	bike	bird	boat	bottle	bus	car
Faster R-CNN[45]	75.8	**82.3**	72.9	56.0	61.2	80.6	85.4
FC-Net	**82.1**	79.1	**74.9**	**61.9**	**63.1**	**81.4**	**85.8**
	cat	chair	cow	table	dog	horse	mbike
Faster R-CNN[45]	83.7	**55.2**	**82.7**	63.2	**82.0**	84.0	80.9
FC-Net	**85.7**	55.1	78.9	**65.6**	**82.0**	**85.6**	**81.0**
	person	plant	sheep	sofa	train	tv	mAP
Faster R-CNN[45]	82.6	**46.5**	73.5	66.8	75.8	**74.2**	73.3
FC-Net	**82.8**	46.4	**75.0**	**72.7**	**80.1**	73.4	**74.6**

图 4.13 自行车类别的激活图

Figure 4.13 The activation maps of bikes.

4.4 本章小结

本章首先提出了基于行人激活图的特征校准与增强遮挡行人检测模块，其中行人激活图是由一些卷积网络提取到的行人视觉模式聚合而形成的。然后我们提出像素级特征校准和目标级区域特征校准，前者通过行人激活图增强了行人可见部分的特征，抑制了遮挡区域特征，从而减少来自遮挡区域的干扰；后者通过局

部搜索，确定自适应的上下文特征提取区域，提取出增强行人边缘的目标级特征。实验结果表明，特征校准网络在行人检测任务上性能有所提升，尤其针对遮挡行人检测尤其有效。最后，我们将它推广至通用性目标检测框架，该方法依然有效。

第5章　基于特征选择-抑制-增强的行人检测

在本章中，我们提出了特征选择-抑制的特征增强方法，它引入了锚点包学习策略，并将目标-特征匹配由手工指定改为动态自适应，实现目标-特征的优化匹配，我们称该算法为多锚点框学习（Multiple Anchor Learning，MAL）。MAL是基于多示例学习[90][91]（Multiple-Instance Learning，MIL）算法来选择优化锚点框，并且对分类和定位进行联合优化。但是，在常规多示例学习方法训练中，这样的迭代选择过程可能难以优化。考虑到在每次学习迭代中选择得分最高的实例可能会产生次优的解[94]，例如得到具有高分类得分但是低定位精度的检测结果，我们提出了特征“选择-抑制-增强”训练机制。这是类似一种对抗的训练方式，它通过扰动得分最高的锚点框的特征来多次降低其置信度，从而让网络考虑更多的锚点框，也就是说更多的低置信度但位置正确的锚点框将有机会参与学习，从而缓解陷入局部最优的情况。这样的求解过程往往会产生一个更佳检测模型。通过将监督方式从独立的锚点框学习转变为多个锚点框学习，MAL 可以充分利用多个锚点框的特征来训练更好的检测器。

我们首先对当前的锚点框匹配和训练策略进行介绍，同时分析锚点框匹配在遮挡问题中的影响。然后提出基于多示例学习的多锚点框学习概念和具体实现。其次，为了联合优化分类和定位两个目标，我们提出特征的“选择-抑制-增强”训练策略，并且详细介绍了实现方式。最后我们验证方法的有效性，呈现实验结果，同时进行分析和总结。

5.1 问题分析

锚点框匹配（Anchor Matching）是目标检测流程中的一个步骤，当前大多数行人检测方法都没有关注锚点框匹配或者样本选择的问题。在图 5.1 中，我们展示了行人的原始标注框和两个候选框，图 5.1（a）展示了数据集标注的信息，其中绿色为行人全身标注框，黄色为行人可见部分标注框。图 5.1（b）在此基础上添加了两个目标候选框（蓝色和红色的候选框）。我们可以看出蓝色的候选框可以覆盖目标，而红色的候选框只覆盖了遮挡部分，它不能提取有效的行人特征，所以蓝色的候选框明显优于红色的候选框。但是在传统锚点框匹配算法中，这两

个候选框都满足正例样本选择标准，即它们和标注框的 IoU 都大于 50%，所以它们在训练中会被无区别对待。

（a）标注框　　（b）候选框示例

图 5.1 遮挡情况下的候选框示例。（a）标注框，绿色为行人全身标注框，黄色为行人可见部分标注框。（b）候选框示例，蓝色和红色边界框为两个候选框。

Figure 5.1 Proposals in the case of occlusion. (a) Ground-truth, green bbox is for the whole body of the pedestrian, and yellow one is for the visible part of the pedestrian. (b) Proposals, blue and red bounding boxes are two proposals.

针对这个问题，文献[52]更新了正例样本的选取准则，即正例样本的选择需要同时满足两个条件。一是行人候选框与标注框的重叠面积的 IoU 需大于阈值α；二是候选框和标注可见框的 IoU 需大于设定的阈值β。该方法针对遮挡行人问题提出了正例样本的选取的新方案。然而，这个方法首先需要额外的标注信息，即可见框的位置，其次这个设计是一种直觉的设计，没有充分考虑匹配过程的学习，参数需要人工设定。

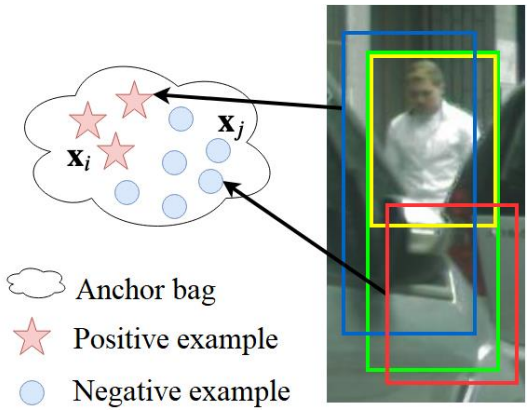


图 5.2 基于多示例学习方法锚点包构建

Figure 5.2 Anchor bag construction based on Multiple-Instance Learning.

除此之外，为了提供丰富的候选框，很多方法在卷积特征上使用手工设计的密集锚点框策略，通过直接的目标-特征匹配进行模型学习。这些锚点框符合均匀分布，并且有固定尺度和宽高比。它们需要确保覆盖各种尺度和外表的目标，这点对于检测器的训练至关重要。合理的锚点框设定可以加速训练的收敛过程，也可以改进检测器的性能。然而，在一组手工设定的锚点框下，并且仅利用空间对齐作（即使用目标与锚点框之间的 IoU）为分配锚点框的标准很难联合优化分类任务和定位任务。如果没有两个优化目标的直接交互，则具有准确定位的检测结果可能具有较低的分类置信度，并受到后端非极大值抑制算法的抑制，如图 5.3（a）基准方法所示。针对该问题，IoU-Net[92]和 FreeAnchor[93]提出了一些解决措施。但是，在训练过程中仍然使用独立的分类和定位置信度。FreeAnchor 根据分类和定位的联合概率选择锚点框，但是考虑到该优化过程的非凸性，基于最大似然估计（Maximum Likelihood Estimation，MLE）的匹配过程并不是最优的。

为此，我们根据多示例学习思想提出锚点包模型（Anchor Bag），在锚点包中可以同时包含正例锚点框和反例锚点框，如图 5.2 所示。在锚点框匹配算法中，使用锚点包代替单独的锚点框优化。这个学习过程并将目标-特征匹配由手工指定改为动态自适应，实现目标-特征的优化匹配。同时，我们还提出使用对抗的机制进行训练，这将产生一个联合优化的结果。

5.2 RetinaNet 回顾

我们提出的多锚点框学习方法选取 RetinaNet[68]作为基准方法，所以本节我们将介绍一下 RetinaNet。RetinaNet 是具有高检测精度的单阶段检测器的代表。由主干网和两个子网组成，一个子网用于目标分类，另一个子网用于目标定位。RetinaNet 主干网络使用特征金字塔网络。根据特征金字塔中的每个特征图，分类子网可以预测类别概率，而回归子网可以使用锚点框作为参考位置来预测目标位置。为了提高效率，两个子网的输入特征在各个特征金字塔层级之间共享。考虑到前景背景类的极端不平衡，也就是在锚点框目标匹配后，正负锚点框损失数值差异较大，需要采用 Focal Loss 来防止训练过程中大量容易的反例样本的损失淹没难样本的损失。

$x \in \mathcal{X}$表示输入图像，$y \in \mathcal{Y}$表示图像的标签。其中$\mathcal{X}$是训练图像集合，$\mathcal{Y}$是

类别的标签集合。定义B是正例图片中的标注框，$b_i \in B$包含类别标签b_i^{cls}和空间位置b_i^{loc}。分类子网络和回归子网络对锚点框a_j分别预测它的分类置信度a_j^{cls}和位置输出a_j^{loc}。如果一幅图像中锚点框的 IoU 和任意标注框的 IoU 大于一定阈值则会被设定为正例锚点框a_{j+}，否则会被设定为反例锚点框a_{j-}。这些锚点框被用于监督网络的学习：

$$\theta^* = \arg\max_{\theta}\left(f_\theta\left(a_{j+}, b_i^{cls}\right) - \gamma f_\theta\left(a_{j-}, b_i^{cls}\right)\right) \tag{5.1}$$

$f_\theta(\cdot)$是分类过程，γ是正反例锚点框重要性的平衡因子。同时，使用正例锚点框来优化目标位置：

$$\theta^* = \arg\max_{\theta} g_\theta\left(a_{j+}, b_i^{loc}\right) \tag{5.2}$$

θ表示网络的参数，$g_\theta(\cdot)$定义为边界框的回归过程。公式 5.1 通过最小化 Focal Loss（$\mathcal{L}_{cls}\left(a_j, b_i^{cls}\right)$）来实现，公式 5.2 通过 Smooth-L1 Loss（$\mathcal{L}_{loc}\left(a_j, b_i^{loc}\right)$）来实现。

在网络学习期间，每个分配的锚点框都独立地监督目标分类和目标定位的学习，而没有考虑锚点框上的分类和定位是否兼容。这可能导致定位精确的锚点框具有较低的分类置信度，这样的锚点框可能会被后端的非极大值抑制过程所丢弃。

5.3 特征选择与抑制

我们提出的特征选择-抑制的特征增强算法是在 RetinaNet 网络结构上进行实现和验证的。该算法通过对分类和定位得分的评估找到最佳的锚点框或者特征选择来更新 RetinaNet。上一节中，我们简要回顾了原始 RetinaNet 的目标分类和定位的机制。接下来，我们将详细阐述 MAL 如何通过评估锚点框来改善分类和定位。最后，我们提出了一种特征“选择-抑制-增强”策略，以寻求 MAL 的最优解。

5.3.1 多锚点框学习概念

分类和回归是目标检测器的两个主要优化目标。在基于卷积神经网络的检测器中，这两个目标通常是在一组固定的候选框或者锚点框上进行优化，获取位置偏差和分类置信度。位置预测准确的锚点框分类置信度不一定高，而分类置信度最大的锚点框位置又不一定准确，这种方式很难真正地对分类和定位进行联合优化。我们提出多锚点框学习（Multiple Anchor Learning，MAL），采用一种多示

例学习方法来选择锚点框并对定位和分类两个模块进行联合优化。MAL 通过构建锚点包，并在其中选择最具有代表性的锚点框进行网络参数优化。与此同时，采用“选择-抑制-增强”的方式，通过扰动锚点框的相应特征来抑制锚点框的置信度，以此来增强分类器的鲁棒性。

在 MAL 的训练阶段，首先计算锚点框与目标边界框之间的交并比，然后选择排序靠前的锚点框构建每个目标的锚点包。MAL 通过结合其分类和定位的分数来评估每个包中的正例锚点框。在每次训练迭代中，MAL 使用选择的正例锚点框来优化训练损失，而不是选择最高得分的锚点框作为最终解。这使得最高分类得分和最高定位得分的锚点框可以同时联合优化，如图 5.3（b）。

（a）基准方法　　（b）MAL

图 5.3 基准方法检测结果与 MAL 检测结果对比。左侧为基准方法，右侧为 MAL。第 1 行是在 NMS 操作之前，第 2 行是最终的检测结果。基准检测器可能会生成具有较低分类分数的高定位精度的边界框（黄色边界框），或具有较高分类分数的低定位精度的边界框（红色边界框），这会导致 NMS 后出现次优结果。MAL 产生的较高分类得分和较好定位的边界框，因此在 NMS 之后可获得更好的检测结果。

Figure 5.3 Detection outputs of the baseline detector (RetinaNet) and the Multiple Anchor Learning (MAL), before and after NMS. The baseline detector may produce bounding boxes with high localization IoU with a low classification score (the yellow bbox), or low localization IoU with a high classification score (the red bbox), which lead to sub-optimal results after NMS. MAL produces bounding boxes with high co-occurrence of top classification and localization, leading to better detection results after NMS.

5.3.2 多锚点框学习损失函数

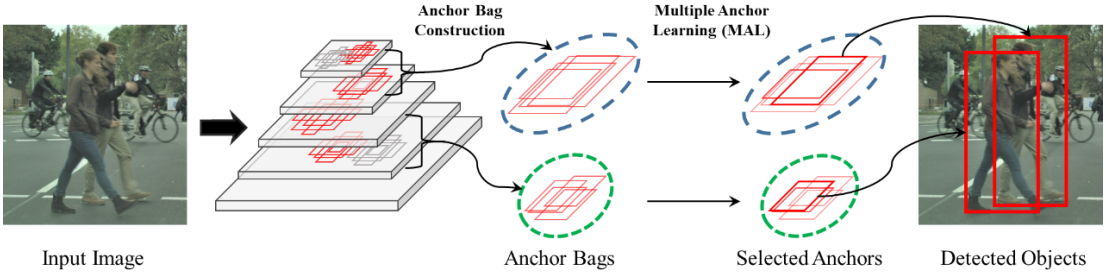


图 5.4 多锚点框学习概念。在特征金字塔中，为每一个目标b_i构建锚点包A_i。MAL 会联合评估每个锚点框分类和定位的置信度，评估出的置信度将用于锚点框的选择，并表示当前锚点框的重要性。

Figure 5.4 The main idea of MAL. In the feature pyramid network, an anchor bag A_i is constructed for each object b_i. Together with the network parameter learning, *i.e.*, back-propagation, MAL evaluates the joint classification and localization confidence of each anchor in A_i. Such confidence is used for anchor selection and indicates the importance of anchors during network parameter evolution.

传统的视觉目标检测算法一般通过直接的目标-特征匹配进行模型学习。然而，针对有些倾斜或遮挡的目标，传统算法并不能差异地看待不同锚点框在训练中的贡献程度。为了缓解单个锚点框优化的缺点，我们构建了锚点包（Anchor Bag），并将目标-特征匹配由手工指定改为动态自适应，实现目标-特征的优化匹配，如图 5.4 所示。在每次学习迭代中，MAL 在锚点包中选择得分较高的实例以更新模型。更新后，模型会以新的置信度评估每个实例，模型学习和锚点框选择将迭代优化目标函数。

为了实现这个目标，我们为第i个目标（标注框）构造一个锚点包A_i。首先计算当前标注框和所有锚点框之间的 IoU，然后选取前k个锚点框构建锚点包。在网络的学习过程中，MAL 将评估A_i中每个锚点框的联合分类和定位置信度。该置信度将用于锚点框的选择，并指示了学习过程中该锚点框的重要程度。为简单起见，仅考虑对正例锚点框的学习，而反例锚点框的学习则仍然采用公式 5.1。MAL 的优化目标函数如下：

$$\begin{aligned}\{\theta^*, a_i^*\} &= \arg\max_{\theta, a_j \in A_i} F_\theta(a_j, b_j) \\ &= \arg\max_{\theta, a_j \in A_i} f_\theta(a_j, b_i^{cls}) + \beta g_\theta(a_j, b_i^{loc})\end{aligned} \quad (5.3)$$

这里$f_\theta(\cdot)$和$g_\theta(\cdot)$分别给出了分类和回归的得分，β是个正则化因子。它旨在为第

i个目标选择最佳的正例锚点框a_i^*，以及学习网络参数θ^*。

将公式 5.2 中定义的目标函数转换为损失函数，如下所示：

$$\begin{aligned}\{\theta^*, \mathrm{a}_i^*\}&=\arg\ min_{\theta,a_j\in A_i}\mathcal{L}_{det}(a_j,b_i)\\&=\arg\ min_{\theta,a_j\in A_i}\mathcal{L}_{cls}(a_j,b_i^{cls})+\beta\mathcal{L}_{reg}(a_j,b_i^{loc})\end{aligned}\quad(5.4)$$

这里$\mathcal{L}_{cls}$和$\mathcal{L}_{reg}$分别是分类和回归的损失。

5.3.3 选择-抑制-增强

优化公式 5.2 或公式 5.4 是非凸问题，这可能会导致锚点框的次优选择。为了缓解该问题并选择最佳锚点框，我们提出通过扰动它们的相应特征来反复降低选定锚点框的置信度。这种学习策略，称为“选择-抑制-增强”优化，以增强特征表示能力，使用这种类似于对抗的方式求解 MAL 问题。

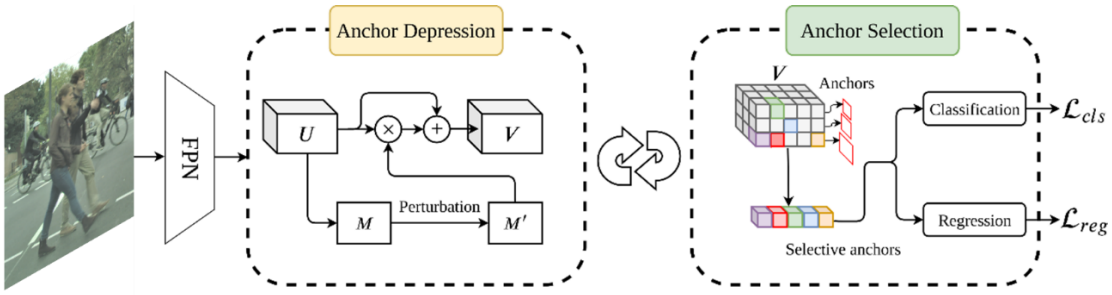


图 5.5 多锚点框学习实现。实现的过程包括锚点框选择和抑制两个模块。U和V分别表示使用锚点框抑制策略之前和之后的卷积特征。M和M`分别表示抑制之前和之后的激活图。

Figure 5.5 MAL implementation. During training, it includes the additional anchor selection and anchor depression modules added to RetinaNet. During test, it uses exactly the same architecture as RetinaNet. U and V respectively denote convolutional feature maps before and after depression. M and M` respectively denote an activation map before and after depression.

5.3.3.1 特征选择

根据$F_\theta(a_j,b_j)$的定义，传统的多示例学习算法倾向于选择得分最高的锚点框。然而，在目标检测的任务上，从每个包中直接选择得分最高的锚点框进行优化，求解将非常困难，这一点已经被连续多示例学习优化方法[94]所验证。在训练阶段，代替公式 5.3 中选择得分最高的锚点框的策略，针对每个锚点包，我们提出了一种 “All-to-Top-1”的锚点框选择策略。当学习迭代过程中，我们线性地减少包中的锚点框的数量从$|A_i|$到 1。我们定义$\lambda=\frac{t}{T}$为变化的进度，其中t是训练中当前迭

代的次数，T是总迭代次数。然后定义$\phi(\lambda)$表示排序靠前的锚点框的索引，并且有 $|\phi(\lambda)| = |A_i| * (1-\lambda) + 1$。最后公式 5.3 被重写成：

$$\{\theta^*, a_i^*\} = \arg max_{\theta, a_j \in A_i} \sum_{j \in \phi(\lambda)} F_\theta(a_j, b_i) \quad (5.5)$$

在早期训练过程中，MAL 利用多个锚点框特征学习检测模型，并在训练最后阶段将收敛于优化单个最佳的锚点框。

5.3.3.2 特征抑制

为了实现对抗机制的训练策略，我们需要降低得分比较高的锚点框的置信度。受到逆注意力网络[95]的启发，我们设计了一种特征抑制过程，通过抖动所选锚点框的特征，以降低其置信度，如图 5.5 所示。该过程是为了将未被选定的锚点框逐步加入到训练过程中，可以使得网络考虑得分最大的锚点框附近的其他候选，这样就完成了对抗的全过程。我们将图像特征和注意力图分别表示为U和M，其中 $M = \sum_l w_l * U_l$，其中w是U特征的权重和，l是U的通道索引。然后，我们通过将高值衰减为零的过程来生成一个新的抑制注意力图$M' = (1 - \mathbf{1}_P) * M$，其中$\mathbf{1}$是 0-1 指示函数。$P$是高数值的位置。通过如下公式进行特征扰动：

$$V = (1 + M') \circ U_l \quad (5.6)$$

这里$\mathbf{1}$是单位矩阵，$\circ$表示逐元素相乘。使用连续优化策略，公式中的抑制过程被写成：

$$V = \left(1 + \left(1 - 1_{\psi(\lambda)}\right) * M\right) \circ U_l \quad (5.7)$$

这里$\psi(\lambda)$表示对多少个像素的值进行抑制，在后续实验中，我们选取了 3 种形式。

5.3.4 优化分析

锚点框的“选择-抑制-增强”策略近似一个对抗的训练过程。首先通过注意力图找到得分最高的锚点框，这些锚点框可将检测损失值 $\mathcal{L}_{det}$降至最低。然后通过抑制操作扰动所选锚点框的相应特征，以降低它们的置信度，这样检测损失值会再次增加。“选择-抑制-增强”策略可帮助学习器找到 MAL 的非凸目标函数的更好解。如图 5.6 的第一条曲线所示，MAL 选择了次优锚点框并陷入了损失函数的局部最小值中。在第二条曲线中，经过锚点框抑制步骤，损失值会再次增加，以致局部最小值被“填充”，因此 MAL 可以继续找到下一个局部最小值。当学习收敛后，MAL 则有更好的机会跳出局部最优，而找到更好的解。

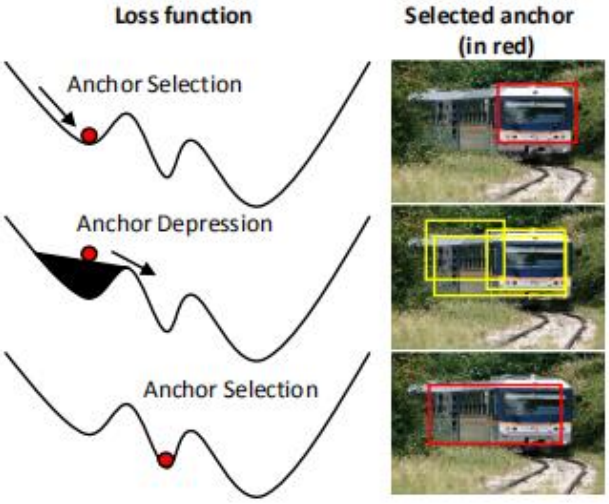


图 5.6 对抗优化过程。在第 1 条曲线中，MAL 选择了次优锚点框并陷入局部最小值。在第 2 条曲线中，锚点框抑制会增加损失值，因此 MAL 会继续优化。这样，MAL 有更大的机会找到最佳的解。

Figure 5.6 Optimization analysis. In the first curve, MAL selects a sub-optimal anchor and gets stuck into a local minimum. In the second curve, anchor depression increases the loss so that MAL continues the optimization. In this way, MAL has a greater chance to find optimal solutions.

5.3.5 方法细节

我们使用带有 FPN 结构的 RetinaNet 检测器作为基准方法，在它的基础上实现了 MAL 算法。锚点框的设置与 RetinaNet 设定相同，共包含 9 种锚点框的设置，其中 3 种锚点框的大小缩放比分别为$\{2^0, 2^{1/3}, 2^{2/3}\}$，3 种锚点框的长宽比$\{1{:}\,2, 1{:}\,1, 2{:}\,1\}$。在各特征金字塔层级上，相对于输入图像尺寸，锚点框可以覆盖从 32 到 813 像素范围。

在网络训练的前馈过程中，我们计算每个锚点框的检测置信度$F_\theta(a_j, b_j)$，以最小化公式 5.4。根据置信度，选择前k个锚点框。然后，在所选锚点框的监督下更新网络参数。选择锚点框之后，再进行锚点框抑制步骤。在下一次迭代中，再次执行锚点框选择步骤以选择高得分的锚点框。

我们方法的推理过程与 RetinaNet 完全相同，通过使用训练好的网络来预测分类分数和目标边界框，然后将其结果送到非极大值抑制算法。由于 MAL 算法仅用于检测器的训练过程，可以学习更有代表性的特征，在测试阶段和基准相同，因此我们学到的检测器可在不增加额外计算成本的情况下实现性能的提升。

5.4 实验结果与分析

为了验证本章所提出的多锚点框学习算法的有效性，我们分别在 ResNet-50、ResNet-101 和 ResNeXt-101[96]网络上进行了实验，并在通用目标数据集 MS-COCO 中测试了该方法。我们进行了消融实验以验证 MAL 中锚点框选择和锚点框抑制的有效性。后面几个小节将分别介绍实验设定、实验分析以及和当前最新方法的对比。

5.4.1 实验设定

我们使用了带有 FPN 的 ResNet-50、ResNet-101 和 ResNeXt-101 作为主干网。批标准化层（Batch Normalization Layer）的参数在训练阶段被固定。每个 GPU 我们使用 2 幅图片作为一个批量，因此在 8 个 GPU 上总共有 16 幅图片进行训练。初始学习率设为 0.01，在训练 ResNet50 的时候，需要训练 135k 次迭代。在 90k 和 120k 训练迭代次数时，学习率衰减 10 倍。而在训练 ResNet-101 和 ResNeXt-101 的时候，需要训练 180k 次迭代，同时在 120k 和 135k 训练迭代次数的时候进行学习率衰减。网络优化使用同步随机梯度下降（Stochastic Gradient Descent，SGD）算法。权重衰减和动量分别设定为 0.0001 和 0.9。在前 500 次迭代时，网络使用线性预热策略（Linear Warmup Strategy）进行训练。我们使用通用目标检测数据集 MS-COCO 进行评测，它包含 80 种目标类别。我们使用约 11.8 万幅图像用于训练，5 千幅图片用于验证，并且约 2 万幅图像用于测试。并用平均准确率（mAP）作为本节的评测指标。在消融实验中，我们使用 ResNet-50 作为基准，并且使用 COCO-minval 数据集用于评测，该集合有 5 千幅测试图片。

5.4.2 特征选择验证实验

表 5.1 锚点包中不同锚点框数量下的检测性能

Table 5.1 Detection performance upon different anchor numbers k in each anchor bag.

Method	AP	AP_{50}	AP_{75}
MAL (k=40)	38.27	56.67	40.81
MAL (k=50)	**38.39**	**56.81**	**41.14**
MAL (k=60)	38.08	56.11	40.18

首先在没有使用抑制模块的前提下，单独验证锚点框选择模块的有效性。我们为锚点包选择不同尺度 k，并且对比了它们的结果，如表 5.1 所示。当 k=40、

50 和 60 时，AP 值变化平稳。当 k=50 时，AP 值取得最大值（38.39%）。在接下来的实验中，我们选取 50 个锚点框用于构建锚点包。表 5.2 展示了不同锚点框的选择训练策略的性能。当使用锚点包训练策略替代原始 RetinaNet 中的分散锚点框设定策略之后，AP 从 35.46%提升到 38.14%，如表 5.2 中 MAL+S (all)所示。在 RetinaNet 中，如果一个锚点框位置准确，但是没有得到最高的分类得分，在损失函数上不会对网络的参数造成影响。而使用锚点包训练策略之后，在检测器训练过程中，这种锚点框会成为一个潜在的候选，从而得到优化。通过使用连续优化策略评测性能进一步提升 AP 至 38.39%，即在训练开始阶段使用所有的锚点框，但是随着训练迭代逐渐减少使用锚点框的数量，如表 5.2 中 MAL+S (all-top1)所示。这验证了连续优化在 MAL 中的有效性，即本章提出的公式 5.5。

表 5.2 锚点框选择策略性能对比。“S”表示选择策略。

Table 5.2 Anchor selection strategy $\phi(\lambda)$. “S” denotes “Selection”

Method	AP	AP_{50}	AP_{75}
RetinaNet	35.46	51.61	39.37
MAL+S (all)	38.14	**56.81**	40.81
MAL+S (all-top1)	**38.39**	**56.81**	**41.14**

5.4.3 特征抑制验证实验

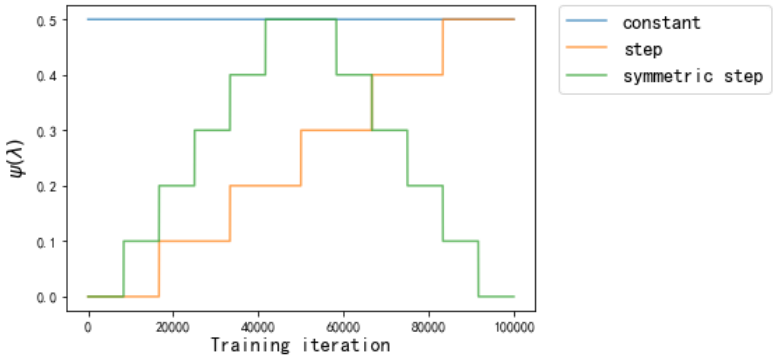


图 5.7 指示函数

Figure 5.7 Different indicator functions

在这一部分实验中，我们仅将抑制模块加入到基准 RetinaNet 中，来找到公式 5.7 中的更合适的指示函数ψ(λ)。我们设计了三种指示函数，分别是常数指示函数（constant）、阶梯指示函数（step）和对称阶梯指示函数（symmetric step），

如图 5.7 所示。第 1 个常数指示函数表示在这个训练过程中保持相同的抑制比例，即在注意力图上，我们抑制前 50%的像素特征。使用这个指示函数之后，AP 值降低了一点，从 35.46%减少到 35.25%，如表 5.3 中 MAL+D (constant)所示。降低的原因可能是在训练开始阶段，网络的参数是随机初始化的，在这个对抗学习过程中抑制模块没有发挥作用。第 2 个阶梯指示函数表示抑制的像素特征选取数量从 0%到 50%呈阶梯状增加，该指示函数对应的性能增长到了 35.88%，如表 5.3 中 MAL+D (step)所示。这表明检测器应该在初始阶段先进行优化，然后再逐渐使用抑制模块。第 3 个指示函数是对称阶梯函数，它呈现一个凸的形状，开始时从 0%增加到 50%，然后再降低到 0%，如图 5.8 中的 symmetric step 所示。在表 5.3 中可以看到，MAL+D（symmetric step）达到了 36.18%的最佳性能。

表 5.3 锚点框抑制策略性能对比。“D”表示抑制策略。常数指示函数、阶梯指示函数和对称阶梯指示函数被比较。

Table 5.3 Depression strategy ψ(λ). “D” denotes “Depression”. The constant function, step function, and symmetric step function are compared.

Method	AP	AP_{50}	AP_{75}
MAL+D (constant)	35.25	51.72	38.92
MAL+D (setp)	35.88	52.34	39.63
MAL+D (symmetric step)	**36.18**	**52.66**	**39.88**

5.4.4 选择-抑制-增强策略分析

首先，我们在特征图注意力图上可视化了 MAL 的影响，如图 5.8 所示。第 1 行和第 3 行是 RetinaNet 的特征图可视化，第 2 行和第 4 行是 MAL 的注意力图可视化。这些注意力图分别是在 10k、50k 和 90k 训练迭代次数时产生的。通过比较发现，MAL 可以激活目标上的更多部分，比如第 1 幅图片中自行车座椅的位置。MAL 也可以抑制背景中的更多区域，比如第 2 幅图片左上角与目标猫无关的红色矩形区域。这表明 MAL 改进了特征，使这些特征更适合于目标检测任务，从而检测器可以更好地检测物体。

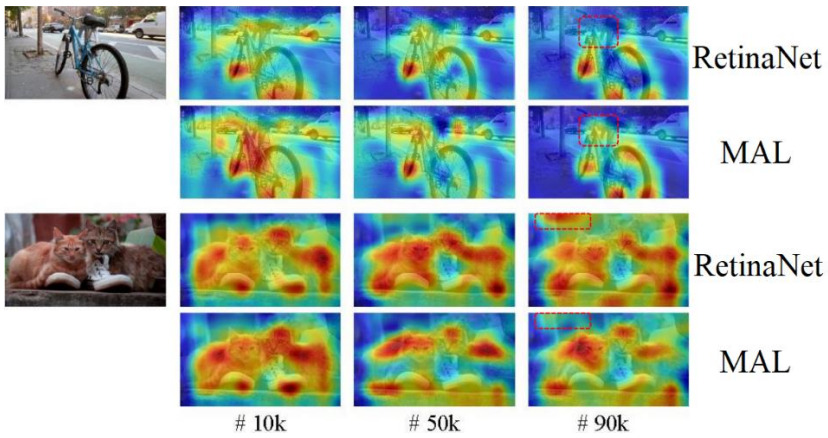


图 5.8 RetinaNet 和 MAL 注意力图对比。这些注意力图是在 10k、50k、和 90k 训练迭代时产生的。如第 90k 次迭代中的红色框所示，MAL 获得了更好的注意力图，从而激活了自行车图像中的更多部分，并抑制了猫图像中的不相关部分。

Figure 5.8 The activation map comparison between RetinaNet (the first and third rows) and MAL (the second and fourth rows). The attention maps at the 10k, 50k and 90k iterations are overlaid on input images. As highlighted by red boxes at the 90k th iteration, MAL gets better attention maps which activate more parts in the bicycle image and suppress irrelevant parts in the cat image.

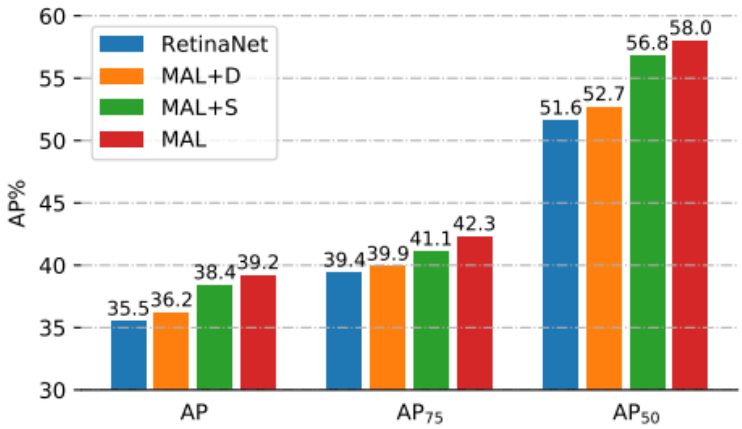


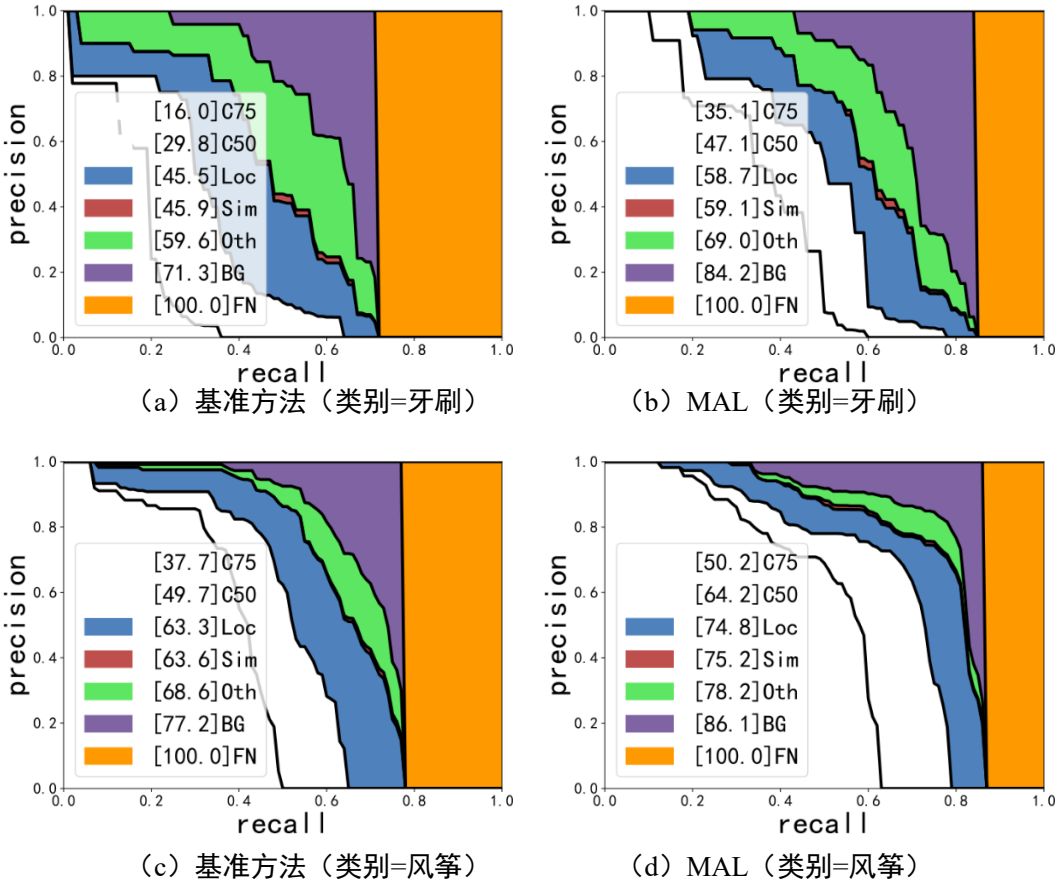
图 5.9 在 COCO-minval 数据集上 MAL 性能对比可视化。在指标 AP，AP_{75} 和 AP^{50} 上，MAL 明显优于基准检测器（RetinaNet）。“S”和“D”分别表示“选择”和“检测”策略。

Figure 5.9 Ablation studies of the anchor selection and depression modules on the COCO-minval dataset. On the metrics AP, AP_{75} and AP_{50}, MAL outperforms the baseline detector (RetinaNet) with significant margins. “S” and “D” respectively denote “Selection” and “Detection”.

在图 5.9 中，我们分析了组合选择和抑制这两个部分的有效性。我们分别比

较了 AP、AP_{50} 和 AP_{75} 这 3 各个指标。原始基准方法的 AP 是 35.5%，在使用抑制模块之后，AP 增加到 36.2%。在使用选择模块之后 AP 进一步增加到 38.4%。当在选择和压抑模块之间采取对抗的方式进行训练时，AP 进一步提高到 39.2%，与原始 RetinaNet 相比，性能提高了 3.7%（35.5% vs. 39.2%）。同时在 AP_{75} 和 AP_{50} 上也呈现了和 AP 相同的增长趋势。

图 5.10 显示了定位结果的误差因素分析。可以看出，较差的定位妨碍了不规则物体（比如倾斜或细长的物体）检测性能的提升。与基准方法相比，我们的方法显著降低了这些目标的定位误差，如图 5.10 中蓝色部分（Loc）。牙刷类别物体的曲线面积（Area Under Curve，AUC）从 15.7%（45.5% − 29.8%）降低到 11.6（58.7% − 47.1%）；对于风筝类别，从 13.6%（63.3% − 49.7%）降低到 10.6%（74.8% − 64.2%）。



（a）基准方法（类别=牙刷）　（b）MAL（类别=牙刷）

（c）基准方法（类别=风筝）　（d）MAL（类别=风筝）

图 5.10 MAL 方法检测性能定性分析

Figure 5.10 Quantitative evaluation of detection performance. Top row: performance comparison on toothbrush detection. Bottom row: performance summary for the kit category.

5.4.5 与其他当前方法进行比较

在消融实验选取最佳参数设定的情况下，我们在表 5.4 中比较了 MAL 和基准 RetinaNet 方法的性能。在使用 ResNet-50 时，MAL 对比基准方法从 35.5%改善到 39.2%，提升了 3.7%。对于 ResNet-101 和 ResNeXt-101，分别改进了 4.5％和 5.1％。这说明了 MAL 在各个主干网上都获得了可靠的提升。

表 5.4 MAL 在 MS-COCO 测试集上和基准方法比较。MAL 显著提升了检测性能。

Table 5.4 Performance comparison with the baseline method (single-scale results) on the MS-COCO test-dev dataset. MAL improves the baseline with significant margins.

Method	Backbone	AP	AP_{50}	AP_{75}
RetinaNet	ResNet-50	35.5	51.6	39.4
MAL (ours)	ResNet-50	**39.2**	**58.0**	**42.3**
RetinaNet	ResNet-101	39.1	59.1	42.3
MAL (ours)	ResNet-101	**43.6**	**62.8**	**47.1**
RetinaNet	ResNeXt-101	40.8	61.1	44.1
MAL (ours)	ResNeXt-101	**45.9**	**65.4**	**49.7**

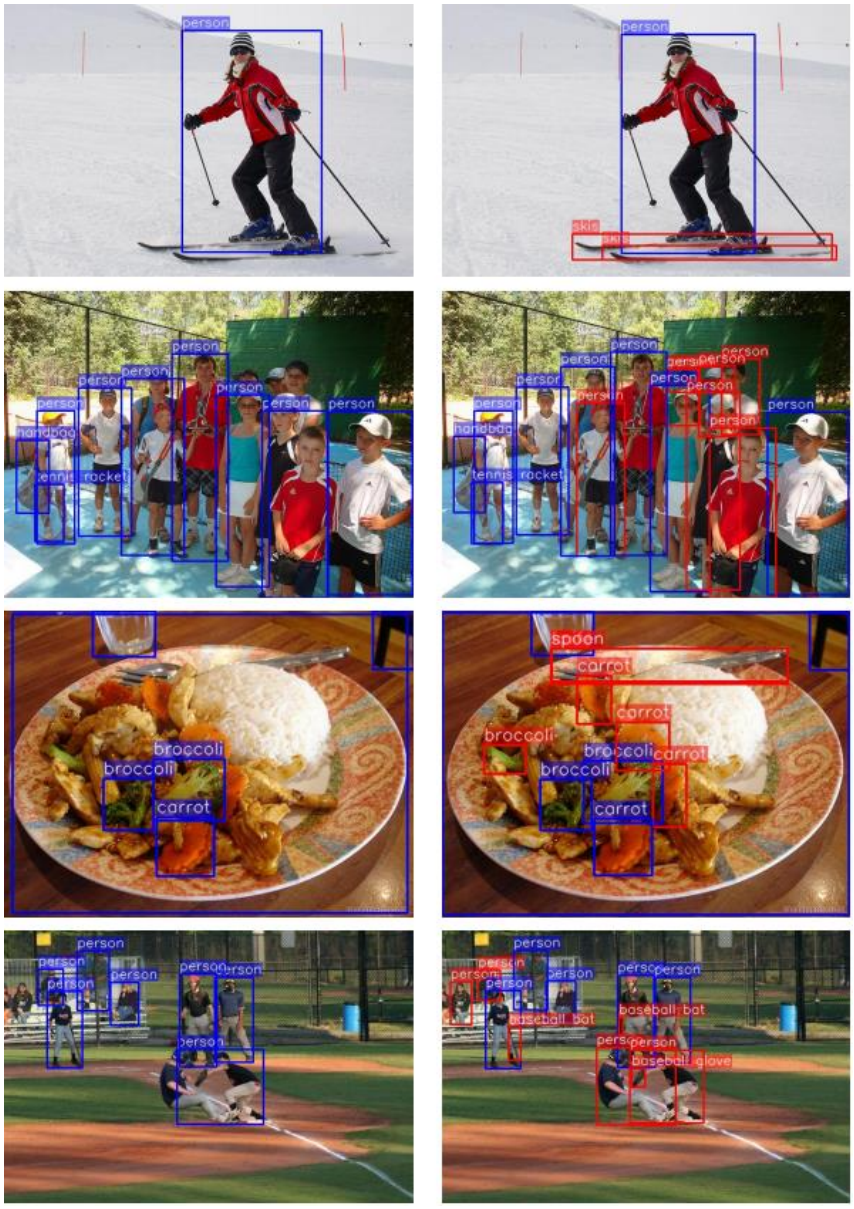
在表 5.5 中，我们在 MS COCO 数据集上比较了 MAL 和单阶段检测器及最新的双阶段检测器，并且按照 AP 的升序进行排列。为了公平的比较，我们将图像进行的缩放，限制图像的短边为 800 像素，并且长边不超过 1333 像素。

对于单阶段检测器，我们选择了最新的 YOLO[67][97]、SSD[66]、FCOS[98]、FreeAnchor[93]和 CenterNet[99]与 MAL 进行对比。使用 ResNet-101 主干网，MAL 可实现单尺度 43.6％的 AP，比无锚点框的方法 FCOS 高 2.1％（43.6％ vs. 41.5％）。使用 ResNeXt-101 主干网，MAL 达到了单尺度的 45.9％的 AP，与最近的 FreeAnchor 相比，可实现 1.1％的增益（45.9 ％ vs. 44.8％）。它还比最新的 CenterNet 高出 1.0％的 AP（45.9％ vs. 44.9％）。但是 CenterNet 使用的是 Hourglass-104[100]作为主干网络，该主干网络具有比 ResNeXt-101 更多的网络参数。针对这种具有挑战性的物体检测任务而言，我们的方法已经具有了足够的优势。在使用 ResNet-101 和 ResNeXt101 的情况下，多尺度测试使 MAL 的 AP 分别提高到 45.0％和 47.0％。

表 5.5 还将 MAL 与具有代表性的双阶段检测器进行了比较，包括带有 FPN[70]的 Faster-RCNN，Mask R-CNN[101]，IoU-Net[92]和 Grid R-CNN[102]。MAL 的检测

性能优于大多数两阶段检测器。特别是，在具有相同主干的情况下，它的性能比最近的 Grid R-CNN 检测器还高出 2.7%（45.9% vs. 43.2%）。作为一种结构较简单的单阶段检测器而言，MAL 已经具有了超越双阶段检测器的巨大潜力。

我们可视化了 MAL 方法检测到的一些结果。可以看出，我们的方法检测检测到了更多细长和遮挡的目标，比如图 5.11 第 1 行右侧的滑雪板和第 3 行右侧的勺子。我们的方法在检测结果可视化上，明显优于基准方法。



(a) RetinaNet (b)MAL

图 5.11 在 MS-COCO 验证集上 RetinaNet 和 MAL 检测结果对比。红色边界框是我们方法相比基准方法成功检测到的目标。

Figure 5.11 Comparison of the detection results between RetinaNet and MAL on the MS-COCO val dataset. Left column: RetinaNet. Right column: MAL.

表 5.5 MAL 在 MS-COCO 测试集上和当前最新方法的性能对比

Table 5.5 Performance comparison with the state-of-the-art methods on the MS-COCO test-dev dataset.

Method	Backbone	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Two-stage method							
Faster R-CNN+++[64]	ResNet-101	34.9	55.7	37.4	15.6	38.7	50.9
Faster R-CNN w FPN[70]	ResNet-101	36.2	59.1	39.0	18.2	39.0	48.2
Faster R-CNN w TDM[103]	Inception-ResNet-v2-TDM	36.8	57.7	39.2	16.2	39.8	52.1
Deformable R-FCN[104]	Inception-ResNet-v2	37.5	58.0	40.8	19.4	40.1	52.5
Mask R-CNN[101]	ResNeXt-101	39.8	62.3	43.4	22.1	43.2	51.2
IoU-Net[92]	ResNet-101	40.6	59.0	-	-	-	-
Cascade RCNN[105]	ResNet-101	42.8	62.1	46.3	23.7	45.5	55.2
Grid R-CNN w/FPN[102]	ResNeXt-101	43.2	63.0	46.6	25.1	46.5	55.2
One-stage methods							
YOLOv2[67]	Darknet-19	21.6	44.0	19.2	5.0	22.4	35.5
SSD513[66]	ResNet-101	31.2	50.4	33.3	10.2	34.5	49.8
YOLOv3[97]	Darknet-53	33.0	57.9	34.4	18.3	35.4	41.9
DSSD513[106]	ResNet-101	33.2	53.3	35.2	13.0	35.4	51.1
GA-RetinaNet[107]	ResNet-50	37.1	56.9	40.0	20.1	40.1	48.0
MetaAnchor[108]	ResNet-50	37.9	-	-	-	-	-
RetinaNet[68]	ResNet-101	39.1	59.1	42.3	21.8	42.7	50.2
CornerNet[109]	Hourglass-104	40.6	56.4	43.2	19.1	42.8	54.3
RetinaNet[68]	ResNeXt-101	40.8	61.1	44.1	24.1	44.2	51.2
FCOS[98]	ResNet-101	41.5	60.7	45.0	24.4	44.8	51.6
FoveaBox[110]	ResNeXt-101	42.1	61.9	45.2	24.9	46.8	55.6
AB+FSAF[111]	ResNeXt-101	42.9	63.8	46.3	26.6	46.2	52.7
FreeAnchor[93]	ResNeXt-101	44.8	64.3	48.4	27.0	47.9	56.0
CenterNet[99]	Hourglass-104	44.9	62.4	48.1	25.6	47.4	57.4
Ours							
MAL	ResNet-101	43.6	62.8	47.1	25.0	46.9	55.8
MAL	ResNeXt-101	45.9	65.4	49.7	27.8	49.1	57.8
MAL (multi-scale)	ResNet-101	45.0	63.7	48.9	28.0	48.0	57.0
MAL (multi-scale)	ResNeXt-101	47.0	66.1	51.2	30.2	50.1	58.9

5.5 本章小结

我们提出了一种有效的锚点框训练方法，称之为多锚点框学习（MAL）。通过锚点框选择机制联合优化检测边界框的分类和定位，MAL 将标准的手工锚点框分配机制升级为可学习的“目标-锚点框”匹配机制。我们还提出了一种简单的“选择-抑制-增强”策略来缓解 MAL 陷入次优化的问题。与基准检测器 RetinaNet 相比，MAL 提升了目标检测的效果。在 MS-COCO 数据集上，与单阶段的检测方法比较中获得了最佳性能，并且优于许多最新的双阶段检测方法。这样的改进不仅考虑了锚点框的最佳选择，而且还考虑了锚点包的隐式的特征装配。我们的工作为锚点框的学习提供了一个新方向。在 MS-COCO 目标检测数据集上的实验表明，MAL 在不同基网络上对比 RetinaNet 均有显著的提升，获得了较好的检测结果。

第6章　总结与展望

行人检测是计算机视觉的一个重要领域，也是场景理解、图像检测、事件检测等许多任务的基础。经过近二十年的深入研究，行人检测技术得到了很大的发展，但是在真实复杂场景中，行人检测算法的性能还有待提升。因此，行人检测也一直是学术界和工业界的研究热点问题。

本文面向图像和视频中的行人和通用目标检测任务，分别从网络结构、特征校准、特征优化三个方面进行了研究，提出了自适应特征增强方法。我们的主要研究包括：环状循环网络用于自适应特征提取，特征校准模块用于遮挡特征增强，选择-抑制-增强的锚点框训练策略优化了特征的学习，将目标-特征匹配准则由手工指定改为动态自适应。

6.1 本文工作总结

本文针对目前行人检测中存在的遮挡问题，提出一些新方法，旨在增强行人和通用目标的特征表示，并且提高在遮挡行人上的检测性能。本文的主要研究成果如下：

(1) 提出一种用于自适应特征增强的网络结构，即环状循环网络（CircleNet），提升了低分辨率行人和遮挡行人检测的准确性。特征金字塔网络是一种层级式的特征提取网络，但是它缺少深层和浅层特征的相互融合。为此，提出环状循环网络（CircleNet）。该网络通过扩展特征金字塔网络结构，增加一条由浅层到深层的通路，并将浅层到深层的通路和深层到浅层的两条通路组合构成环状结构，通过权值共享可以将它视为环状循环网络。该网络通过往复式的特征适配，提取更有表达性的行人特征，该特征在保持高分辨率的同时具有更强的语义信息。实验中验证了环状循环网络多次对图像中的行人进行检测的有效性。同时结合行人实例分解训练策略，使得环状循环网络的潜力得到了进一步发挥，在一般行人和遮挡行人上提升了检测的准确性。

(2) 提出特征校准（FC）模块，增强了行人可见部分特征，同时抑制了来自遮挡区域的噪声干扰。这是一种解决遮挡行人检测的特征增强方法，首先我们定义了行人激活模式概念，该模式是深度卷积特征每个通道学到的行人局部特征，

比如脚步、手臂和头部等。然后聚合所有激活模式，即可得到行人激活图。使用该激活图可加强图像级别的特征，进行像素级特征校准。在提取目标级特征之后，我们再使用目标级区域特征校准增强特征，该模块融合自适应的上下文信息，可以学习背景和行人的共生信息。我们的方法可以适应性地根据遮挡情况加强或减弱特征，最终使得模型可以检测不同遮挡程度的行人，在不同遮挡率的情况下都可以鲁棒地检测行人。

(3) 提出特征选择-抑制的特征增强方法，称之为多锚点框学习（MAL）算法。传统的锚点框学习算法中，正例锚点框的选择只考虑了锚点框和标注框之间的 IoU，即通过直接的目标-特征匹配进行学习。然而，针对有些倾斜或遮挡的目标，传统的方法不能差异地看待不同锚点框在训练中的贡献程度。因此，我们提出了基于多示例学习的多锚点框学习算法，该算法引入锚点包学习策略，并将目标-特征匹配准则由手工指定改为动态自适应，实现目标-特征的优化匹配。除此之外，为了联合优化目标定位和目标分类，我们提出使用特征“选择-抑制-增强”的对抗训练策略，这缓解了在优化锚点框时陷入到局部最优的情况。实验中，在选取的 3 个主干网下，MAL 方法比所选取的基准方法得到的性能都有所提高。

6.2 未来工作展望

行人检测仍然是一项充满了挑战的研究课题。尽管本文针对其中的几个关键问题进行了一些探索和尝试，取得了一些研究成果。但是，必须指出的是行人检测仍然是一个开放问题，将现有技术直接应用于无人驾驶场景等，在应用上还存在一些明显不足，有待于研究和完善。结合本文研究中所遇到的问题，关于未来的研究工作提出以下展望：

(1) 循环中信息过滤与选择算法：本文提出的循环特征增强网络虽然解决了行人不同分辨率和遮挡情况特征提取的问题，但仍然面临信息过滤与选择的困扰。在有效循环中信息被多次加工，结合门控函数实现信息的选择和丢弃，将进一步扩展循环特征增强网络的能力。

(2) 行人稳定区域的非极大值抑制算法：本文针对遮挡行人检测，提出了行人自激活图。行人激活图可以指示行人可见区域，让网络具有对遮挡的感知能力。对比行人激活图，我们也尝试了使用一个分割子网络预测行人可见部分中心区域，通过弱监督语义分割的训练策略，该网络可以生成行人稳定的可见部分区域。这

种改进已经验证了有效性，可以被进一步探索。由于非极大值抑制后处理阶段丢失了一些拥挤场景中正确的检测结果，我们可以利用行人可见部分的稳定区域代替行人的边界框作为非极大值抑制算法的输入，改进检测器在拥挤场景下的性能。稳定区域非极大值抑制算法，为拥挤行人检测和密集通用目标检测提供了全新的解决思路。

(3) 推广至通用目标检测：本文提出的方法都可以推广至通用目标检测。循环特征增强网络对通用目标检测中的遮挡目标和弱小目标上也存在优势。针对特征校准模块，在实验中我们已经在通用目标检测中进行了探索。通过平均各个类别的激活图可以得到前景激活图，再由前景激活图加强通用目标特征。在未来的工作中，可以结合网络学习能力为每个类别单独生成该类激活图的权重，更好的各类激活图的融合方式需要被进一步所探索。将这种特征加强方法推广至通用目标检测，在实际应用中有着非常重要的意义。

参考文献

[1] Silver D, Schrittwieser J, Simonyan K, et al. Mastering the game of go without human knowledge[J]. Nature, 2017, 550(7676): 354-359.

[2] Silver D, Huang A, Maddison C J, et al. Mastering the game of Go with deep neural networks and tree search[J]. Nature, 2016, 529(7587): 484-489.

[3] Vinyals O, Ewalds T, Bartunov S, et al. Starcraft ii: A new challenge for reinforcement learning[J]. arXiv preprint arXiv:1708.04782, 2017.

[4] Haugeland J. Artificial intelligence: The very idea[M]. MIT press, 1989.

[5] 李彦冬. 基于卷积神经网络的计算机视觉关键技术研究[D]. 电子科技大学, 2017.

[6] 费驰. 多尺度遮挡鲁棒的全天候行人检测技术研究[D]. 中国科学技术大学, 2019.

[7] 王爱丽. 基于计算机视觉的行人交通信息智能检测理论和关键技术研究[D]. 北京交通大学, 2016.

[8] 中国人民共和国中央人民政府, 国务院关于印发新一代人工智能发展规划的通知. http://www.gov.cn/zhengce/content/2017-07/20/content_5211996.htm

[9] Szeliski R. Computer vision: algorithms and applications[M]. Springer Science & Business Media, 2010.

[10] LeCun Y, Bengio Y, Hinton G. Deep learning[J]. Nature, 2015, 521(7553): 436-444.

[11] 李媛媛. 基于深度学习的肺结节检测研究[D]. 电子科技大学, 2019.

[12] 丁鹏. 基于深度卷积神经网络的光学遥感目标检测技术研究[D]. 中国科学院大学 (中国科学院长春光学精密机械与物理研究所), 2019.

[13] 王文玉. 海量天文光谱数据中白矮主序双星的发现研究[D]. 济南: 山东大学, 2015.

[14] Dollar P, Wojek C, Schiele B, et al. Pedestrian detection: An evaluation of the state of the art[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012, 34(4): 743-761.

[15] 张苗辉. 基于视觉系统的行人检测与跟踪方法研究[D]. 上海交通大学, 2013.

[16] 盛碧云. 基于特征学习的人体目标检测和分析[D]. 东南大学, 2017.

[17] 付新川. 图像中的行人检测关键技术研究[D]. 电子科技大学, 2019.

[18] 田广. 基于视觉的行人检测和跟踪技术的研究[D]. 上海交通大学, 2007.

[19] 王蓦. 视频监控中的行人检测与再识别研究[D]. 武汉大学, 2016.

[20] 郭萍. 基于视频的人体行为分析[D]. 北京: 北京交通大学, 2012.

[21] 连旭. 基于步态的身份识别研究与实现[D]. 辽宁科技大学, 2014.

[22] Chen C, Seff A, Kornhauser A, et al. Deepdriving: Learning affordance for direct perception in autonomous driving[C]. Proceedings of the IEEE International Conference on Computer Vision. 2015: 2722-2730.

[23] Smail H, David H, Larry S D. W4: Real-Time surveillance of People and their Activities[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000, 22(8).

[24] Ye Q, Zhang T, Ke W, et al. Self-learning scene-specific pedestrian detectors using a progressive latent model[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 509-518.

[25] Gerónimo D, López A M. Vision-based pedestrian protection systems for intelligent vehicles[M]. New York, NY, USA: Springer, 2014.

[26] Geiger A, Lenz P, Stiller C, et al. Vision meets robotics: The KITTI dataset[J]. The International Journal of Robotics Research, 2013, 32(11): 1231-1237.

[27] 国家统计局交通事故死亡人数. http://data.stats.gov.cn/

[28] 刘弋锋. 基于浅层学习引导深度学习的行人检测[D]. 武汉大学, 2016.

[29] Vehicles A. 3.0: Preparing for the Future of Transportation[J]. Federal Policy Framework. National Highway Transportation Safety Administration, US Department of Transportation, 2018.

[30] Viola P, Jones M J, Snow D. Detecting pedestrians using patterns of motion and appearance[J]. International Journal of Computer Vision, 2005, 63(2): 153-161.

[31] Dalal N, Triggs B. Histograms of oriented gradients for human detection[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2005, 1: 886-893.

[32] 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, 2010, 32(9): 1627-1645.

[33] Dollár P, Appel R, Belongie S, et al. Fast feature pyramids for object detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014, 36(8): 1532-1545.

[34] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[C]. Advances in Neural Information Processing Systems. 2012: 1097-1105.

[35] Oren M, Papageorgiou C, Sinha P, et al. Pedestrian detection using wavelet templates[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 1997, 97: 193-199.

[36] Levi K, Weiss Y. Learning object detection from a small number of examples: the importance of good features[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2004, 2: II-II.

[37] Ojala T, Pietikainen M, Maenpaa T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002, 24(7): 971-987.

[38] Hinterstoisser S, Lepetit V, Ilic S, et al. Dominant orientation templates for real-time detection of texture-less objects[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2010: 2257-2264.

[39] Watanabe T, Ito S, Yokoi K. Co-occurrence histograms of oriented gradients for pedestrian

detection[C]. Pacific-Rim Symposium on Image and Video Technology. Springer, Berlin, Heidelberg, 2009: 37-47.

[40] Tuzel O, Porikli F, Meer P. Region covariance: A fast descriptor for detection and classification[C]. Proceedings of the European Conference on Computer Vision. 2006: 589-600.

[41] Hosang J, Omran M, Benenson R, et al. Taking a deeper look at pedestrians[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 4073-4082.

[42] Tian Y, Luo P, Wang X, et al. Pedestrian detection aided by deep learning semantic tasks[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 5079-5087.

[43] Girshick R. Fast r-cnn[C]. Proceedings of the IEEE International Conference on Computer Vision. 2015: 1440-1448.

[44] Zhang L, Lin L, Liang X, et al. Is faster r-cnn doing well for pedestrian detection?[C]. Proceedings of the European Conference on Computer Vision. 2016: 443-457.

[45] Ren S, He K, Girshick R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks[C]. Advances in Neural Information Processing Systems. 2015: 91-99.

[46] Zhang S, Benenson R, Schiele B. Citypersons: A diverse dataset for pedestrian detection[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 3213-3221.

[47] Kingma D P, Ba J. Adam: A method for stochastic optimization[J]. arXiv preprint arXiv:1412.6980, 2014.

[48] Wang S, Cheng J, Liu H, et al. Pcn: Part and context information for pedestrian detection with cnns[J]. arXiv preprint arXiv:1804.04483, 2018.

[49] Song T, Sun L, Xie D, et al. Small-scale pedestrian detection based on somatic topology localization and temporal feature aggregation[J]. arXiv preprint arXiv:1807.01438, 2018.

[50] Liu W, Liao S, Ren W, et al. High-level Semantic Feature Detection: A New Perspective for Pedestrian Detection[J]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019.

[51] Mathias M, Benenson R, Timofte R, et al. Handling occlusions with franken-classifiers[C]. Proceedings of the IEEE International Conference on Computer Vision. 2013: 1505-1512.

[52] Zhou C, Yuan J. Bi-box regression for pedestrian detection and occlusion estimation[C]. Proceedings of the European Conference on Computer Vision. 2018: 135-151.

[53] Zhang S, Wen L, Bian X, et al. Occlusion-aware R-CNN: detecting pedestrians in a crowd[C]. Proceedings of the European Conference on Computer Vision. 2018: 637-653.

[54] Brazil G, Yin X, Liu X. Illuminating pedestrians via simultaneous detection & segmentation[C]. Proceedings of the IEEE International Conference on Computer Vision. 2017: 4950-4959.

[55] Zhou C, Wu M, Lam S K. SSA-CNN: Semantic Self-Attention CNN for Pedestrian

Detection[J]. arXiv preprint arXiv:1902.09080, 2019.

[56] Zhang S, Yang J, Schiele B. Occluded pedestrian detection through guided attention in CNNs[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 6995-7003.

[57] Lin C, Lu J, Wang G, et al. Graininess-aware deep feature learning for pedestrian detection[C]. Proceedings of the European Conference on Computer Vision. 2018: 732-747.

[58] Liu S, Huang D, Wang Y. Adaptive NMS: Refining Pedestrian Detection in a Crowd[J]. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019.

[59] Wang X, Xiao T, Jiang Y, et al. Repulsion loss: Detecting pedestrians in a crowd[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 7774-7783.

[60] Uijlings J R R, Van De Sande K E A, Gevers T, et al. Selective search for object recognition[J]. International Journal of Computer Vision, 2013, 104(2): 154-171.

[61] Zitnick C L, Dollár P. Edge boxes: Locating object proposals from edges[C]. Proceedings of the European Conference on Computer Vision. 2014: 391-405.

[62] Arbeláez P, Pont-Tuset J, Barron J T, et al. Multiscale combinatorial grouping[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2014: 328-335.

[63] 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.

[64] 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.

[65] 曹俊豪. 基于深度学习的行人检测算法研究[D]. 北京邮电大学, 2019.

[66] Liu W, Anguelov D, Erhan D, et al. SSD: Single shot multibox detector[C]. Proceedings of the European Conference on Computer Vision. 2016: 21-37.

[67] 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.

[68] Lin T Y, Goyal P, Girshick R, et al. Focal loss for dense object detection[C]. Proceedings of the IEEE International Conference on Computer Vision. 2017: 2980-2988.

[69] Chen L C, Papandreou G, Kokkinos I, et al. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 40(4): 834-848.

[70] Lin T Y, Dollár P, Girshick R, et al. Feature pyramid networks for object detection[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 2117-2125.

[71] Cordts M, Omran M, Ramos S, et al. The cityscapes dataset for semantic urban scene understanding[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern

Recognition. 2016: 3213-3223.
[72] Dollár P, Wojek C, Schiele B, et al. Pedestrian detection: A benchmark[J]. 2009.

[73] Geiger A, Lenz P, Urtasun R. Are we ready for autonomous driving? the kitti vision benchmark suite[C]. 2012 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2012: 3354-3361.

[74] Everingham M, Van Gool L, Williams C K I, et al. The pascal visual object classes (voc) challenge[J]. International Journal of Computer Vision, 2010, 88(2): 303-338.

[75] Shao S, Zhao Z, Li B, et al. Crowdhuman: A benchmark for detecting human in a crowd[J]. arXiv preprint arXiv:1805.00123, 2018.

[76] 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.

[77] Van der Maaten L, Hinton G. Visualizing non-metric similarities in multiple maps[J]. Machine learning, 2012, 87(1): 33-55.

[78] Van Der Maaten L. Accelerating t-SNE using tree-based algorithms[J]. The Journal of Machine Learning Research, 2014, 15(1): 3221-3245.

[79] Khoreva A, Benenson R, Hosang J, et al. Simple does it: Weakly supervised instance and semantic segmentation[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 876-885.

[80] Brazil G, Yin X, Liu X. Illuminating pedestrians via simultaneous detection & segmentation[C]. Proceedings of the IEEE International Conference on Computer Vision. 2017: 4950-4959.

[81] Tian Y, Luo P, Wang X, et al. Deep learning strong parts for pedestrian detection[C]. Proceedings of the IEEE International Conference on Computer Vision. 2015: 1904-1912.

[82] Cai Z, Fan Q, Feris R S, et al. A unified multi-scale deep convolutional neural network for fast object detection[C]. Proceedings of the European Conference on Computer Vision. 2016: 354-370.

[83] Pang Y, Xie J, Khan M H, et al. Mask-guided attention network for occluded pedestrian detection[C]. Proceedings of the IEEE International Conference on Computer Vision. 2019: 4967-4975.

[84] Zhou B, Khosla A, Lapedriza A, et al. Learning deep features for discriminative localization[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 2921-2929.

[85] Zhang S, Benenson R, Omran M, et al. How far are we from solving pedestrian detection?[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016:

1259-1267.

[86] Hu J, Shen L, Sun G. Squeeze-and-excitation networks[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 7132-7141.

[87] Zagoruyko S, Lerer A, Lin T Y, et al. A multipath network for object detection[J]. arXiv preprint arXiv:1604.02135, 2016.

[88] Wang H, Li Y, Wang S. Fast Pedestrian Detection With Attention-Enhanced Multi-Scale RPN and Soft-Cascaded Decision Trees[J]. IEEE Transactions on Intelligent Transportation Systems, 2019.

[89] CityPersons 官方网站，https://bitbucket.org/shanshanzhang/citypersons/src/default/.

[90] Andrews S, Tsochantaridis I, Hofmann T. Support vector machines for multiple-instance learning[J]. Advances in Neural Information Processing Systems, 2003: 577-584.

[91] Ramazan G., Jakob V., Cordelia S. Weakly Supervised Object Localization with Multi-fold Multiple Instance Learning. IEEE Conference on Computer Vision and Pattern Recognition. 2014: 2409-2416.

[92] Jiang B, Luo R, Mao J, et al. Acquisition of localization confidence for accurate object detection[C]. Proceedings of the European Conference on Computer Vision. 2018: 784-799.

[93] Zhang X, Wan F, Liu C, et al. Freeanchor: Learning to match anchors for visual object detection[C]. Advances in Neural Information Processing Systems. 2019: 147-155.

[94] Wan F, Liu C, Ke W, et al. C-MIL: Continuation multiple instance learning for weakly supervised object detection[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019: 2199-2208.

[95] Huang Z, Ke W, Huang D. Improving object detection with inverted attention[C]. 2020 IEEE Winter Conference on Applications of Computer Vision. IEEE, 2020: 1294-1302.

[96] Xie S, Girshick R, Dollár P, et al. Aggregated residual transformations for deep neural networks[C]. Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 1492-1500.

[97] Redmon J, Farhadi A. YOLO9000: better, faster, stronger[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 7263-7271.

[98] Tian Z, Shen C, Chen H, et al. Fcos: Fully convolutional one-stage object detection[C]. Proceedings of the IEEE International Conference on Computer Vision. 2019: 9627-9636.

[99] Duan K, Bai S, Xie L, et al. Centernet: Object detection with keypoint triplets[J]. arXiv preprint arXiv:1904.08189, 2019, 1(2): 4.

[100] Newell A, Yang K, Deng J. Stacked hourglass networks for human pose estimation[C]. Proceedings of the European Conference on Computer Vision. 2016: 483-499.

[101] He K, Gkioxari G, Dollár P, et al. Mask r-cnn[C]. Proceedings of the IEEE International Conference on Computer Vision. 2017: 2961-2969.

[102] Lu X, Li B, Yue Y, et al. Grid r-cnn[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019: 7363-7372.

[103] Shrivastava A, Sukthankar R, Malik J, et al. Beyond skip connections: Top-down modulation for object detection[J]. arXiv preprint arXiv:1612.06851, 2016.

[104] Dai J, Qi H, Xiong Y, et al. Deformable convolutional networks[C]. Proceedings of the IEEE International Conference on Computer Vision. 2017: 764-773.

[105] Cai Z, Vasconcelos N. Cascade r-cnn: Delving into high quality object detection[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 6154-6162.

[106] Fu C Y, Liu W, Ranga A, et al. Dssd: Deconvolutional single shot detector[J]. arXiv preprint arXiv:1701.06659, 2017.

[107] Wang J, Chen K, Yang S, et al. Region proposal by guided anchoring[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019: 2965-2974.

[108] Yang T, Zhang X, Li Z, et al. Metaanchor: Learning to detect objects with customized anchors[C]. Advances in Neural Information Processing Systems. 2018: 320-330.

[109] Law H, Deng J. Cornernet: Detecting objects as paired keypoints[C]. Proceedings of the European Conference on Computer Vision. 2018: 734-750.

[110] Kong T, Sun F, Liu H, et al. FoveaBox: Beyond Anchor-based Object Detector. arXiv 2019[J]. arXiv preprint arXiv:1904.03797.

[111] Zhu C, He Y, Savvides M. Feature selective anchor-free module for single-shot object detection[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019: 840-849.

附录 中英文对照表

人工智能	Artificial Intelligence，AI
计算机视觉	Computer Vision，CV
深度学习	Deep Learning，DL
先进驾驶辅助系统	Advanced Driver Assistance Systems, ADAS
行人保护系统	Pedestrian Protection Systems, PPSs
行人检测系统	Pedestrian Detection System，PDS
梯度方向直方图特征	Histograms of Oriented Gradient Features，HOG
可变行组件模型	Deformable Parts Model，DPM
积分通道特征	Integral Channel Feature, ICF
深度卷积神经网络	Deep Convolutional Neural Network, DCNN
局部二值化特征	Local Binary Pattern Features，LBP
主导方向模板特征	Dominant Orientation Template Features，DOT
共现特征	Co-Occurrence Features
协方差特征	Covariance Features
标注框	Ground-truth
锚点框	Anchor
候选框	Proposal
候选区域提取网络	Region Proposal Network，RPN
交并比	Intersection over Union，IoU
单位图像下的虚警率	False Positives Per Images，FPPI
平均对数丢失率	Log-average Miss Rate，MR
平均精确率	Mean Average Precision，mAP
随机梯度下降	Stochastic Gradient Descent，SGD

非极大值抑制	Non-Maximum Suppression, NMS
自激活	Self-activation，SA
行人激活图	Pedestrian Activation Map，PAM
特征校准	Feature Calibration, FC
特征金字塔网络	Feature Pyramid Network, FPN
全局池化层	Global Average Pooling, GAP
多锚点框学习	Multiple Anchor Learning, MAL
锚点框匹配	Anchor Matching
最大似然估计	Maximum Likelihood Estimation，MLE
线性预热策略	Linear Warmup Strategy

致　谢

雁栖湖畔，岁月如梭。转眼间博士学习生涯却已接近尾声，非常有幸在这所优秀的学府度过美好的时光，并完成我的博士学业。回首望去，少许遗憾，同时也有奔向新旅程的兴奋。在雁栖湖畔，我收获了珍贵的友谊、美好的时光和宝贵的知识。在此毕业论文完成之际，我想由衷地向曾经给予我无数帮助的老师、同学、朋友和家人表示深深的谢意！

首先，我要特别感谢我的恩师叶齐祥教授！叶老师严谨细心的科研精神让我钦佩，广阔的学术视野和学术造诣让我受益匪浅。在学术科研上他对我悉心指导，在论文发表过程中他为我进行细心的评阅和修改，在此表示由衷地感谢。他除了教会我如何科研，还教会了我如何思考。此外，还要感谢叶老师在生活上对我的关心和照顾，让我远离家乡时也能感受到如亲人般的关照。

其次，衷心感谢实验室焦建彬教授、韩振军副教授和秦飞副教授。焦老师为人谦和、治学严谨，为我们提供了良好的科研环境和科研指导，是我强大的后盾。韩老师风趣幽默，在科研和学习中对我给予很大的帮助。秦老师思维缜密，在我论文的写作过程中给予过很多指导。

我也要由衷地感谢华为诺亚方舟实验室的刘健庄老师，在华为诺亚方舟实验室实习的日子是我博士期间宝贵的回忆，刘老师治学严谨求实，一丝不苟，让我受益良多。感谢伯克利大学的许慧娟老师，她在繁忙的科研中还抽时间和我进行实验细节的讨论，我从这些讨论中获得了很多启发。

感谢师兄师姐、师弟师妹——柯炜、李策、陈孝罡、魏朋旭、高山、邹佳凌、张晓丹、崔妍婷、庞丽金、黄显淞、刘畅、周彦钊、朱艺、王攀、薛昊岚、张小松、余学辉、姚远、苗彩敬和所有 PriSDL 实验室的同学们。感谢你们在科研上的帮助与合作，感谢一起学习生活中结下的深厚友谊。感谢同届好友李兆举、戴蔚群、万方、王忻雷在学习生活中对我的帮助。

最后要感谢我的父母和我的未婚妻崔千对我的默默支持和无私奉献，你们是我坚强的后盾，是我前进的动力，是我克服一切苦难的信心来源。

张天亮

2020 年 8 月

作者简历及攻读学位期间发表的学术论文与研究成果

作者简历：

2009 年 09 月—2013 年 06 月，在武汉理工大学信息工程学院获得学士学位。
2013 年 09 月—2017 年 06 月，在中国科学院大学工程科学学院获得硕士学位。
2017 年 09 月—2020 年 08 月，在中国科学院大学电子电气与通信工程学院攻读博士学位。

已发表（或正式接受）的学术论文：

[1] **Tianliang Zhang**, Zhenjun Han, Huijuan Xu, Baochang Zhang, Qixiang Ye. CircleNet: Reciprocating Feature Adaptation for Robust Pedestrian Detection[J]. *IEEE Transactions on Intelligent Transportation Systems (ITS)*, 2019. **(中科院一区国际期刊, 影响因子 6.319)**

[2] Wei Ke, **Tianliang Zhang (equal contribution)**, Zeyi Huang, Qixiang Ye, Jianzhuang Liu, Dong Huang. Multiple Anchor Learning for Visual Object Detection. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.（**CCF A类会议，共同一作**）

[3] Qixiang Ye, **Tianliang Zhang**, Wei Ke. Progressive Latent Models for Self-learning Scene-specific Pedestrian Detectors [J]. *IEEE Transactions on Intelligent Transportation Systems (ITS)*, 2019. **(中科院一区国际期刊, 影响因子 6.319，导师一作)**

[4] Qixiang Ye, **Tianliang Zhang**, Wei Ke, Qiang Qiu, Jie Chen, Guillermo Sapiro, Baochang Zhang. Self-learning Scene-specific Pedestrian Detectors using a Progressive Latent Model. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. （**CCF A类会议，导师一作**）

已撰写的论文:

[1] **Tianliang Zhang**, Qixiang Ye, Baochang Zhang, Jianzhuang Liu, Xiaopeng Zhang, Qi Tian. Feature Calibration Network for Pedestrian Detection in the Wild [J]. *IEEE Transactions on Intelligent Transportation Systems (ITS) (Major Revision)*.

申请或已获得的专利：

[1] 叶齐祥，**张天亮**，刘健庄，张晓鹏，田奇，江立辉，行人检测方案、装置、计算机可读存储介质和芯片，201910697411.3，中国发明型专利，已受理。

参加的研究项目及获奖情况：

[1] 国家自然科学基金重点项目，视觉目标自学习建模与在线处理，2019 年 1 月-2023 年 12 月。

[2] 华为技术有限公司横向课题，基于弱监督学习的视觉目标检测，2020 年 1 月-2021 年 12 月。

[3] 博士生国家奖学金，2019 年。

[4] 中国科学院大学，三好学生， 2017~2018 学年。