中国科学院大学
University of Chinese Academy of Sciences

博士学位论文

行人目标特征表示与检测方法研究

作者姓名：柯　炜

指导教师：叶齐祥　教授

中国科学院大学电子电气与通信工程学院

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

2018 年 2 月

Feature Representation and Pedestrian Detection

A dissertation/thesis 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

Wei Ke

Supervisor： Professor Qixiang Ye

School of Electronic, Electrical and Communication Engineering

University of Chinese Academy of Sciences

June 2016

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：

日　　期：

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：　　导师签名：

日　　期：　　日　　期：

摘 要

行人检测是计算机视觉的重要任务之一，其目标是判别一幅图像或者视频序列中是否存在行人并给出精确位置。由于行人检测关注对象是人这类特殊而重要的目标，而且还可以为图像视频检索、目标跟踪和图像分类等其它计算机视觉相关领域提供支持，因此具有重要研究意义。行人检测可应用于智能视频监控、车辆辅助驾驶以及智能机器人等领域，具有实际应用价值。

行人检测中，目标表示和目标定位是其检测框架中最主要的两个部分。表示能力强的中层特征能够增强检测模型的判别性，提高行人检测的性能。描述精确的底层特征能够用来获得行人候选区域，在全监督行人检测中可以降低检测时间，在无监督行人检测中可以降低样本搜索空间。本文针对行人检测的底层和中层特征表示，以及其在全监督和自学习行人检测器中的应用进行研究，主要工作如下：

(1) 提取了基于侧输出残差网络(Side-output Residual Network, SRN)的行人底层特征提取方法。该方法使用残差单元(Residual Unit, RU)拟合残差单元的输出与真实值之间的误差。通过自深到浅依次堆叠残差单元，侧输出残差网络以拟合多尺度上的误差流替代直接拟合输出。该方法不仅可以抑制复杂背景，还可以有效地选择对称轴或边缘的尺度。将其扩展成多分支侧输出残差网络结构后，可以同时提取输入图像的对称性信息和边缘信息等底层特征。

(2) 提出了基于贝叶斯得分重排序(Baysesian Scoring based Proposal Reranking)的行人候选区域提取方法。传统方法采用超像素合并的候选区域提取方式虽然能够实现精确定位，但由于无法计算置信度而存在大量冗余；采用置信度的候选区域提取方式虽然定位不精确但是可以排序。本文根据这两类方法的互补性提出了基于贝叶斯得分重排序的候选区域提取方式。对于一幅输入图像，通过多分支侧输出网络残差同时获得边缘响应图和对称性响应图，并且使用这两个响应图在贝叶斯框架下计算超像素合并产生的冗余区域的得分。选取得分高的候选区域子集在保证了召回率的前提下，降低了冗余度。

(3) 提出了基于 PCA 卷积特征(PCA Filters Based Convolutional Channel

Features, PCA-CCF)的全监督行人检测方法。在特征设计上借鉴了卷积神经网络的优点，不同的是使用 PCA 得到简化的卷积核，对聚合通道特征进行张量运算。该特征不仅通过 PCA 滤波器的正交性对聚合通道特征去相关，而且增加了更多的特征通道，以此增强表达能力。在检测框架上借鉴了 R-CNN 的思想，通过弱分类器获得一些候选区域之后，采用 PCA-CCF 并结合级联 AdaBosst 分类器进行精细分类。实验表明，PCA-CCF 不依赖于深度学习框架确能够有效地提高行人检测精度。

(4) 合作提出了基于渐进优化模型(Progressive Latent Model, PLM)的自学习行人检测方法。全监督行人检测中，需要大量的标注样本，工作量庞大。对于特定场景的监控视频，自学习行人检测器通过渐进优化方式，迭代地进行目标发现、目标增强和标签传播，进而达到不使用任何标注样本而仅通过自动学习的方式得到行人检测器的目的。相比于传统的隐模型方法，这种渐进优化模型增加了空间约束项，在降低候选目标搜索空间的同时加强了目标定位的准确性，同时，使用基于图模型的标签传播算法获得更多的正例样本以及难反例样本，增加了分类器的多样性与判别性。

关键词：行人检测，特征学习，深度学习，候选区域提取，自学习分类器

Abstract

Pedestrian detection is one of the most important branches in computer vision, the goal of which is to determine whether there is pedestrian in one image or video sequence, and if such exists, locate the pedestrian. Pedestrian detection focuses on human, the most important object on the earth, and is useful for other computer vision tasks, such as retrieval, tracking, and classification. Pedestrian detection is widely applied in many practical systems including intelligent video surveillance, driver assistant system, intelligent robotics, etc.

In pedestrian detection framework, the two most import parts are feature representation and object location. With discriminative middle-level features, the detection performance can be increased in supervised pedestrian detection. With accurate low-level features, pedestrian candidates are obtained, which increases the computational efficiency of the supervised methods and decreases the search space of samples in self-learning method. This dissertation conducts research on the low- and middle- level feature representation for supervised and self-learning pedestrian detection, and the contributions of this dissertation are as follows:

(1) A Side-output Residual Network (SRN) is proposed to extract low-level features for pedestrian images. SRN leverages output Residual Units (RUs) to fit the errors between the object symmetry ground truth and the outputs of RUs. By stacking RUs in a deep-to-shallow manner, SRN exploits the ‘flow’ of errors among multiple scales to ease the problems of fitting complex outputs with limited layers, suppressing the complex backgrounds, and effectively matching symmetry or edge of different scales. The proposed SRN is further updated to a multi-task side-output residual network for joint symmetry and edge detection.

(2) A Bayesian Scoring based Proposal Reranking is proposed to release pedestrian candidates. Although classical segmentation-based object proposal approach can produce region proposals with high localization accuracy, it also incorporates

significant redundancy because of the lack of object confidence used to evaluate the proposals. The objectness-based approach gives the confidence without localization accuracy. With the complementary of the two kinds of approaches, we propose Bayesian Scoring based Proposal Reranking to release pedestrian candidates. For an image, the contour and symmetry are extracted by Multi-task Side-output Residual Network and used to score the bounding box with a Bayesian framework. A subset of high-scored proposals can not only guarantee the recall rate, but also decrease the redundancy significantly.

(3) A kind of PCA filters based convolutional channel features (PCA-CCF) for supervised pedestrian detection. For pedestrian representation, we use the convolutional network architecture with pre-learned PCA filters to enhance the aggregate channel features (ACF). PCA-CCF reduces the correlation among ACF with the orthogonal PCA filters and increases the representation ability by more feature channels. Besides, we use an R-CNN like detection framework. On the candidates generated by weak classifier, PCA-CCF features and cascaded AdaBoost classifier are used for fine classification. Experiments show that PCA-CCF has significant pedestrian detection performance gain compared with ACF.

(4) A Progressive Latent Model (PLM) is proposed for self-learning pedestrian detection. In supervised way, pedestrian detection involves lots of human annotation. However, the self-learning approach, for scene-specific surveillance video, is deployed as progressive steps of object discovery, object enforcement, and label propagation and the pedestrian detector is learned without any data annotation involved. Compared with conventional latent models, the proposed PLM incorporates a spatial regularization term to reduce ambiguities in object proposals and to enforce object localization, and a graph-based label propagation to discover more positive and hard instances in adjacent frames to increase the diversity and discriminability.

Key Words：Pedestrian Detection, Feature Learning, Deep Learning, Object Proposal, Self-learning

目 录

图目录

表目录

第 1 章 绪论

1.1 引言

人们每天都在通过视觉、听觉、嗅觉、触觉和味觉感受着外界传来的纷繁复杂的信息。进入信息时代之后，视觉、多媒体等信息的比重高达 70%以上，而且还在不断地增加，处理这些信息需要大量的复杂运算。计算机视觉目标检测技术可以帮助我们自动从这些庞大的图像集合或者视频中提取出感兴趣的物体，与此同时还为图像视频检索、目标跟踪和图像分类等其它计算机视觉相关领域提供技术支撑。作为计算机视觉领域的重要和关键研究内容，视觉目标检测在过去几十年中取得了重要的进展并且仍然是领域研究热点问题之一，历久弥新。

在各类视觉目标中，行人是最有代表性的一类目标，行人目标检测是应用领域最广泛的一项关键技术。在智能视频监控中，行人检测可以告诉我们视频是否有行人，位置在哪儿，帮助人们更加直观、准确和及时地获得监控区域的相关信息；在交通辅助驾驶系统中，行人检测可以预警驾驶环境复杂路况，防止交通意外的发生；智能家居中，各种智能系统都是以行人为服务的主体，行人检测是必不可缺的一部分。

1.2 课题研究背景与意义

1.2.1 课题的来源

本文研究工作受到国家重点基础研究发展计划（973 计划）、国家自然科学基金重点课题及两项自然科学基金面上项目资助，研究成果在飞行器进近威胁目标检测、智能监控系统中得到了初步验证。

1、“基于多源数据的飞行器进近威胁目标检测跟踪及行为预测”，国家自然科学基金重点项目（课题编号：61039003），2011.01-2014.12，已结题。

2、“飞行器威胁目标识别与图像鲁棒匹配理论与方法”，国家 973 计划子课题（课题编号：2010CB731804-2），2010.01-2014.12，已结题。

3、“多视角多姿态人体目标检测”，国家自然科学基金面上项目，（课题编号：

61271433），2013.01-2016.12，已结题。

4、“弱监督视觉目标检测”，国家自然科学基金面上课题，(课题编号：61671427)，2017.01-2020.12，在研。

1.2.2 课题的背景

近十几年来，虽然涌现出大量的行人检测以及目标检测的方法，但检测的性能却仍然受到诸多限制，制约着实际应用。这些影响因素主要包括：(1) 大部分目标非刚性。在目标检测中我们更关注那些可以自由移动的物体，尤其是行人这类目标，在移动时往往是非刚性多姿态的。(2) 受自然环境的影响。在不同的自然环境中，行人目标呈现出不同的形式。交通场景中，行人存在高光照的部分，亮度出现饱和；行人处在阴影中的部分对比度又会比较低。这些都为中层特征增加了难度。(3) 多视角现象。横看成岭侧成峰，从不同视角看待同一行人目标会产生不同的形状， 为目标的表示带来挑战。(4) 遮挡问题。复杂的监控和交通场景中，行人容易互相遮挡，更容易被其它的物体遮挡。遮挡会导致行人物体信息的丢失，增加检测难度。 (5) 复杂的背景也给行人检测产生了很大的干扰。一些行人检测示例图片如图 1-1 所示，一幅图像中往往同时存在多种因素。

为了鲁棒、快速地进行行人检测，研究者们提出了各种各样的手工设计特征(Hand-craft Features)以更好地对行人进行表示，使其在特征空间中能够有效地和其他物体区别开来。这些手工设计的特征计算简单，使得在特征图像特征金字塔上逐层逐像素扫窗(Sliding Window)的方式成为在待检测图像或视频中进行行人定位的流行方式。近几年出现的深度学习特征(Deep Learning Feature)，从某种程度上解决了非刚性、光照、阴影、多视角等问题，使其表示能力远大于纯手工设

图 1-1 行人图像数据示例

计的特征。然而，由于深度学习特征受到需要的样本量大、计算复杂度高、计算耗时长的制约，扫窗这一框架因候选区域过多而被舍弃，逐步兴起的是一些基于分割或者弱分类器的行人候选区域提取方法。这些方法虽然降低了待检测的窗口的个数，但基于深度学习提取特征的目标检测方法的计算时间复杂度依然很高。

深度学习自主地学习网络中各个节点的参数，而且网络越复杂往往可以带来检测性能的进一步提升，但同时检测所需时间也进一步加大。受到手工设计特征的启发，如果在特征学习网络中加入一些人工设计（先验）来简化网络结构，特征提取时间就会减少。与此同时，优化候选区域提取的方式以进一步减少待检测窗口个数，也会达到减少目标检测所需时间目的。因此，设计更好的行人底层和中层特征表示对于提高行人检测的速度和精度都具有重要的研究价值。

1.2.3 课题的应用领域

行人检测涉及到特征提取、数字图像与视频处理、模式识别以及机器学习等多个方面，是典型的计算机视觉问题。同时，由于其关注对象是人，因此是在目标检测领域被格外关注的一类研究问题。行人检测以图像或者视频为目标，回答是否包含行人和行人在哪儿两个问题，其典型应用有如下几个方面：

(1) 智能视频监控

智能视频监控中，最关心的就是行人这类目标。在道路十字路口，可以有效地分析行人的活动，保障公共区域安全；在小区，可以发现可疑人员，降低安全风险；在机场、火车站以及景点，可以通过检测行人判定客流量，及时地采取必要的应急措施；在天网系统中，行人检测可以大大降低可疑人员的检索空间。有效的行人检测，在高效地保证公共场所的安全的同时，可以大量节省人力物力资源，产生庞大的社会价值。

(2) 交通辅助驾驶

汽车已经成为现代人出行不可或缺的一个重要工具，但每日各种交通事故都给人们带来很大威胁。在交通环境中如果能有效地进行行人报警，使驾驶员有更多的反应和处理时间，可以大大降低交通事故的发生。谷歌、百度等互联网及人工智能公司及各大汽车厂商目前均致力于自动驾驶研究及应用，为行人检测技术提供了广阔的应用空间。同时，行人目标检测还能够为机场跑道清理、危险预警

图 1-2 行人检测的应用领域

等应用提供支撑。

(3) 智能家居

智能家居的服务对象是人，因此，更加准确的行人检测技术能够让机器人更好地为人服务，让人机交互更加友善。

(4) 其它计算机视觉任务

行人检测也是很多计算机任务的基础，包括跟踪以及轨迹分析、行人再识别等。现有的很多跟踪和行人再识别系统都还未真实的考虑到检测问题，而是手工给定一个初始值。随着技术的发展，自动检测与跟踪、检测与行人再识别这类更加接近实际应用的系统会受到越来越多的关注。

1.3 行人检测中的研究问题

作为目标检测的一种，行人检测有着目标检测所共有的一些难点，同时，也存在本领域的一些特有问题。作为目标检测的一种，行人检测不仅可以和目标检测一样进行全监督的训练，也可以使用先验信息在不进行标注的情况下训练自学习分类器。它也受到 R-CNN 检测框架的冲击，是继续保持原有的扫窗框架，还是采用新的候选区域之后精细分类的框架。作为特殊的目标，行人检测中又有着自己特征提取的独特方式。如何改进检测框、设计新特征、减少监督信息（数据标注）都是行人检测中可研究的问题。

减少监督：目标检测从开始的人脸检测[1]到行人检测[2]，从多目标检测[3]到细粒度检测[4]，基本都是采用全监督的方式。一方面在早期，硬件的计算能力并不是很强大，能够处理的数据有限。对于有限的数据，进行目标的标定工作量并

不是很大，因此数据集不大时基本都可以做到完备的标注[2, 3]。数据量对标注的影响也体现在最近流行的 ImageNet[5]和 COCO[6]目标检测竞赛，其目标检测所使用的标注数据量要远小于图像分类的数据量。另一方面，全监督算法发展得更为完善，无论是支持向量机(SVM, Surpport Vector Machine) [7]，贝叶斯分类器[8]、Boost 方法[9]还是随机森林[10]，都已经有了很完备的体系支撑，做目标检测这类计算机视觉任务的研究者需要把更多的精力放在特征设计上。

硬件计算能力的增加为处理大量数据带来可能，但数据量的增大又会提高全监督学习的标注成本，因此，各种从算法层面或者从问题定义方面来解决弱监督或者无监督的工作大量涌现。算法层面，迁移学习(Transfer Learning)[11]或者多实例学习(MIL, Multiple Instance Learning)[12]是常用的方法。从问题定义方面，在多目标检测中，可只给定图像标注行弱监督检测，这种标注的代价并不大。而由于行人这类目标的特殊性，甚至可以使用无监督的方式训练自学习分类器。其可行性在于行人的先验信息特别多，也在于行人视频数据容易采集。

框架选择：目标检测早期研究中最成功的一种方法是 Viola 和 Jones 提出的级联 Adaboost 检测器[1]。他们的方法中的很多设计思想被延续到行人检测当中，包括多尺度图像金字塔扫窗，以及高效率的积分图特征计算。

多尺度金字塔扫窗有三方面的优势。第一个优势是由扫窗的固有特性决定的。扫窗是指在一幅特征图像上，用固定大小的窗口，逐像素自左向右，自上至下依次扫动，获得窗口内的特征，然后对该窗口进行分类。这种逐像素操作，能够最大限度的遍历整个图像。很多时候有些分类得分很高的窗口重叠很多，这种现象也可以通过非极大值抑制得到解决。对于固定尺度的目标，这种方法的召回率无疑是百分之百。但是，如果考虑到目标的大小并不固定，就需要采用不同大小的窗口进行扫窗，窗口中特征维度不能得到有效的对齐，而这恰恰是多尺度金字塔扫窗所带来的后两个优势：天然的多尺度和特征对齐方式。构建多尺度特征金字塔可以用同样大小的窗口扫窗：在大尺度特征图上扫窗相当于检测小一些的目标，而在小尺度特征图上扫窗相当于检测大一些的目标。用同样大小的窗口，其特征的维度自然相同。在多目标检测中，即使使用多尺度金字塔扫窗，仍然不能很好地解决目标长宽比不固定的问题。然而，对于行人这类特殊目标，因为站立的行

人宽高比比基本都在 0.5 附近，因此直接取 0.5 就可以适应大多数行人样本。

多尺度金字塔扫描虽然具有以上优势，但由于需要逐像素扫描，该扫窗方式中需要判定的窗口数量在数百万级别，同时具有很高的冗余性；另一方面，当特征太复杂时，多尺度金字塔的构建几乎是不可能的。

在深度目标检测框架 R-CNN 出现之前，多目标检测的结果是远远低于行人这种单目标的检测结果的。然而，随着 R-CNN[13]在多目标检测领域取得巨大成功，行人检测也就面临是否要采用这种新框架的问题。在 R-CNN 中，首先通过无监督的候选区域提取方式，提取大约两千个候选区域，然后使用表示能力大大提升的深度学习特征对每个窗口进行特征提取并进行分类。候选区域是针对广义的目标而设计的，因此其提取过程中直接包含了多尺度多长宽比的目标区域，解决了多尺度金字塔扫窗的长宽比问题，但其召回率却不如后者容易控制。由于候选区域数量比多尺度金字塔扫窗中的数量降低了数个数量级，因此深度学习这种复杂的特征也就能够顺利得被利用。

R-CNN 在做多目标检测时，在候选区域的召回率和分类器性能上达到了比较好的均衡，使得多目标检测性能大幅提升。然而，在行人检测中，即使使用传统特征仍能达到不错的性能，而且针对广义的目标目标设计的候选区域提取方法在行人检测上的召回率并不理想。这些问题在很长一段时间内限制了 R-CNN 在行人检测上的应用。

特征设计：VJ[1]中的另一个被广泛应用的设计是积分图特征，如图 1-3 所示。积分图在很长一段时间都为各种手工设计特征提供了平台。积分图特征的特点是对于任何一个窗口的特征只需要通过该窗口的四个拐角点的值进行计算。在图 1-3 中，假如点 1 处存储的是整个 A 区域的特征向量，点 2 存储 A 和 B 两个区域的特征，点 3 存储 A 和 C 两个区域的特征，点 4 存储 A、B、C 和 D 区域的特征，那么获得 D 区域的特征向量只需对四个点进行简单的运算就可以得到。这种像素点位置存储整个区域的像素特征的图被称为积分图。只要特征是针对像素的，积分图就可以被有效地使用起来，在很大程度上降低特征提取的速度。由于手工特征都是针对像素设计的，因此基本都可以使用积分图进行加速。

提取各种各样的手工设计特征，并使用积分图加速后做特征提取在很长一段

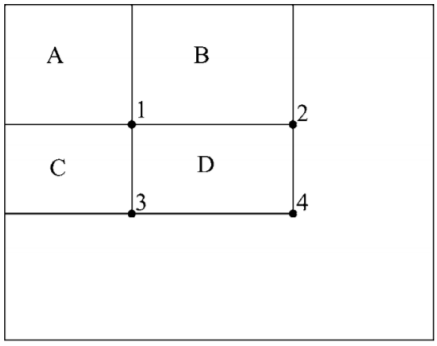


图 1-3 特征积分图示意图

时间内都是行人检测的主要研究内容。但是，手工设计特征的描述能力和区分性并不是很强。近几年来，采用学习的方式获得更强描述能力的特征成为趋势。

随着 2006 年 Hinton 在科学杂志上发表的神经网络的论文[14]和他的研究组在 2012 年 ImageNet 图像分类竞赛中取得巨大成功[5]，深度学习特征的应用进入快速增长状态。深度学习特征具有强大的特征描述能力，但是其提取计算代价非常大，很难像传统特征一样做多个尺度，而且往往也是针对一整幅图像提取特征，也很难像手工设计特征那样通过积分图快速提取某个区域的表达。除此之外，行人检测中将行人样本先缩放再提特征也不是很好的方式。因此深度学习特征在行人检测中并没有完全跟上多目标检测，很久一段时间都没有被有效得利用起来[15]。

1.4 研究内容与主要贡献

本文研究内容及其关系如图 1-4 所示。对于行人检测而言，全监督方式长期作为研究对象[16]，而近期自学习方法也被触及[17]，但类似于多目标的弱监督方式只出现过短暂的热潮。对于全监督行人检测，同时结合卷积特征和手工设计特征优势，本文提出了基于 PCA 的卷积特征[16]。在[16]中，候选区域的提取仍然采用传统由粗到精(Coarse to Fine)的策略，使用弱分类器提供候选区域，因此其有更高的召回率。为了提高流行的候选区域提取方式的召回率，本文一方面提出了基于贝叶斯得分的无监督候选区域提取方法[18,19]，另一方面设计了更强的底层特征[20-22]。这些无监督的候选区域提取方法又可以用来降低自学习行人检测器的搜索空间[17]。本文的研究内容与主要贡献有以下几个方面：

(1) 提取了基于侧输出残差网络(Side-output Residual Network, SRN)的行人

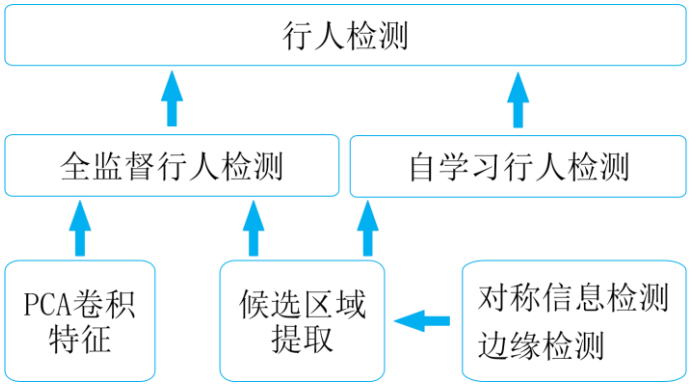


图 1-4 研究内容关系图

底层特征提取方法。该方法使用残差单元(Residual Unit, RU)拟合残差单元的输出与真实值之间的误差。通过自深到浅依次堆叠残差单元，侧输出残差网络利用拟合多尺度上的误差流替代了直接拟合输出，不仅可以抑制复杂背景，并且可以有效地选择对称轴或边缘的尺度。将其扩展成多分支侧输出残差网络结构后，可以同时提取输入图像的对称性信息和边缘信息。

(2) 提出了基于贝叶斯得分重排序(Baysesian Scoring based Proposal Reranking)的行人候选区域提取方法。采用超像素合并的候选区域提取方式定位精确，但由于无法计算置信度而存在大量冗余；采用置信度的候选区域提取方式定位不精确但是有序。本文根据这两类方法的互补性提出了基于贝叶斯得分重排序的候选区域提取方式。对于一幅输入图像，通过多分支侧输出网络残差同时获得边缘响应图和对称性响应图，并且被用来在贝叶斯框架下计算超像素合并产生的冗余区域的得分。选取得分高的候选区域子集在保证了召回率的前提下，降低了冗余度。

(3) 提出了基于 PCA 卷积特征(PCA Filters Based Convolutional Channel Features, PCA-CCF)的全监督行人检测方法。在特征设计上借鉴了卷积神经网络，但是使用 PCA 得到固定的卷积核，对聚合通道特征进行张量运算。该特征不仅通过 PCA 滤波器的正交性对聚合通道特征去相关，而且增加了更多的特征通道，以此增强表达能力。在检测框架上借鉴了 R-CNN 的思想，通过弱分类器获得一些候选区域之后，采用 PCA-CCF 并结合级联 AdaBosst 分类器进行精细分类。实验表明，PCA-CCF 能够有效地提高聚合通道特征的行人检测精度。

(4) （合作）提出了基于渐进优化(Progressive Latent Model, PLM)的自学习行人检测方法。全监督行人检测中，需要大量的标注样本，工作量庞大。对于特定场景的监控视频，自学习行人检测器通过渐进优化方式，迭代地进行目标发现、目标增强和标签传播，进而达到不使用任何标注样本自动学习得到行人检测器的目的。相比于传统的隐模型方法，这种渐进优化模型增加了空间约束项，在降低候选目标搜索空间的同时加强了目标定位的准确性，同时，使用基于图模型的标签传播算法获得更多的正例样本以及难反例样本，增加了分类器的多样性与判别性。

1.5 本文组织结构

第一章，绪论。论述了行人检测的研究背景与研究意义，分析了当前行人检测中存在的难点和常见问题，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，行人检测研究的发展与现状。介绍目标检测框架的更迭到先进行候选区域提取再进行精细检测的新框架及其原因，总结分析了现有的候选区域提取方式的分类，对全监督行人检测和自学习行人检测进行了详细的整理和分析。

第三章，基于侧输出网络的行人底层特征提取。首先描述了深度学习网络在图像底层特征提取中的发展情况，然后详细介绍了侧输出残差网络以及多分支侧输出残差网络的设计原理，最后，在多个公开数据集上与多种方法进行了边缘检测和对称性检测的性能比较与分析。

第四章，基于贝叶斯得分重排序的候选区域提取。首先论述了典型的两种无监督候选区域提取方法的互补性，然后介绍了超像素合并方式产生冗余候选区域的相似性自适应搜索，以及利用底层边缘特征和对称性特征构建贝叶斯得分计算给定窗口包含目标的置信度并对冗余候选区域重排序。在多个数据集上，通过与现有的候选区域提取方法的比较，验证了基于贝叶斯得分重排序的候选区域提取方法的有效性。

第五章，基于 PCA 卷积特征的全监督行人检测。提出基于 PCA 卷积特征并将其用来描述行人。通过 PCA 卷积增加特征描述能力的同时，对特征进行了正

交投影。由于特征计算复杂度增加，采用弱分类器获得一些候选区域，再使用 PCA 卷积特征进行精细分类，以降低分类时间。实验中对比了全监督行人检测中候选区域提取的最佳方式，以及通过行人检测来验证 PCA 卷积特征的作用。

第六章，基于渐进增强的自学习行人检测器。首先介绍了自学习行人检测器的框架，然后详细介绍了自学习检测器中的行人发现、行人增强以及标签传播算法，最后比较和分析了自学习检测与迁移学习、弱监督以及全监督方法的在特定场景监控视频中的检测性能。

第七章，总结了本文的主要内容，并对未来工作进行了展望，包括采用深度学习进一步提高底层特征的提取能力、改进端到端的全监督行人检测以及合理地在自学习分类器中进行使用等。

第 2 章 行人检测的发展与现状

现阶段的目标检测，主要分为数据集驱动的多目标检测和任务驱动的单目标检测。在单目标检测中，与日常生活最相关的人脸、行人和车辆是研究机构和各大公司最关注的目标。相比于人脸检测中基本都是裁切的正面形态数据和车辆检测中的刚性目标车体，行人检测因其非刚性导致的姿态多样、视角多样，更加具有挑战性。行人检测的方法可以推广到多类目标检测中。例如在行人检测中取得巨大成功的梯度方向直方图特征(HOG, Histogram of Oriented Gradients)[2]，被扩展到多目标检测中，并发展成为了里程碑式的工作 DPM(Deformable Part Model)[23]。多目标检测的工作也为行人检测提供思路，例如多目标检测的另一个里程碑工作 R-CNN[13]为行人检测框架提供了一些新的借鉴。

本章首先介绍了目标检测框架从扫窗到 R-CNN 的发展，再介绍了新框架中不可或缺的候选区域提取相关工作，最后介绍行人检测的发展，包括全监督行人检测也包括自学习行人检测两部分。

2.1 目标检测框架更迭

视觉目标检测从本质上说是由目标表示和目标定位两部分构成，目标表示要回答目标是什么，目标定位要回答目标在哪儿，其框架如图 2-1 所示。而目标表示和目标定位两个阶段要处理的最关键问题分别是特征提取和候选区域提取。训练阶段，将待检测的目标作为正例样本，背景作为反例样本，按照一定的特征

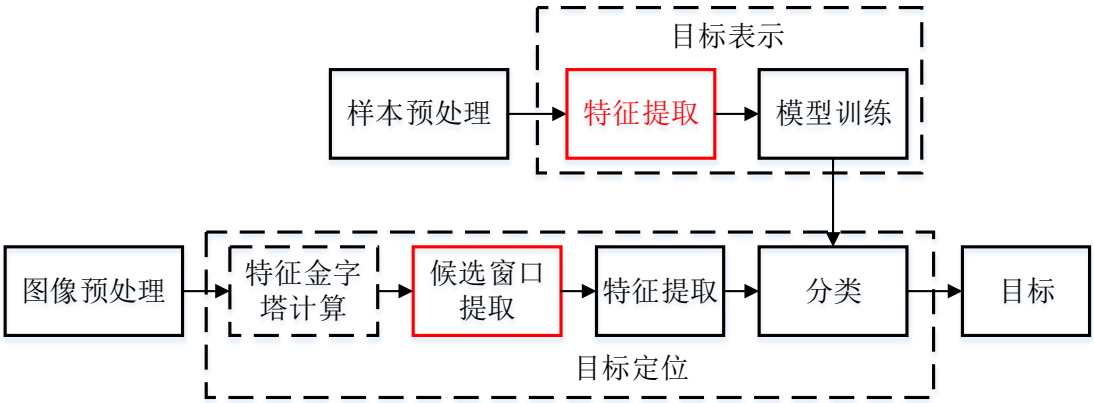


图 2-1 目标检测的流程

提取方法将目标从图像空间映射到特征空间中，利用 SVM 或者 Boost 等分类算法，训练二分类模型，将正例和反例样本分开。通过提取特征和训练模型就可以完成目标表示，知道什么是待检测目标。测试阶段，对于一幅新输入图像，在图像中获得候选窗口并将其映射到特征空间之后，利用已经训练的模型确定哪个候选窗口为待检测目标，达到目标定位的目的。等分类算法，训练二分类模型，将正例和反例样本分开。通过提取特征和训练模型就可以完成目标表示，知道什么是待检测目标。测试阶段，对于一幅新输入图像，在图像中获得候选窗口并将其映射到特征空间之后，利用已经训练的模型确定哪个候选窗口为待检测目标，从而实现目标定位。

目标检测框架更迭主要来自于测试阶段流程的不同。传统扫窗方法中，候选区域是通过多尺度逐像素移动同样大小的窗口获得窗口。虽然这种方法获得窗口数量巨大，但其每个窗口仍然可以被看做一个候选区域窗口。由于数量太多，所以需要采用多尺度特征金字塔将测试图像特征提取完成，采用查表方式对每个候选区域进行特征提取。对于新的 R-CNN 框架，由于其特征描述提取太过复杂，并不能轻松构建多尺度特征金字塔，其采用对每个候选区域单独提取特征的方式，即没有图 2-1 中虚线框的特征金字塔计算部分。

新的框架在多目标检测中取得了很大的成功[9, 24-28]，检测性能得到了大幅度提升。由于无监督的候选区域提取方式在全监督行人检测中并不理想，因此滞后于多目标检测数年，这种新框架才在行人检测中获得了成功的应用[29]。本文关注于新框架在全监督行人检测的应用，一方面尝试解决候选区域提取的改进，希望其能成功应用在行人检测中，另一方面沿着传统解决方法的思路设计具有更强表达能力的特征。

2.2 候选区域提取

传统的目标定位方法通过在多尺度图像特征金字塔上逐像素扫描窗口，并用分类器对每个窗口的特征向量进行分类，确定其是目标物体还是背景而达到定位的目的，如图 2-2 所示。在扫窗框架中，为了减少每个窗口计算特征所需要的时间，需要首先在图像金字塔上提取特征构建特征金字塔，然后在图像金字塔每层

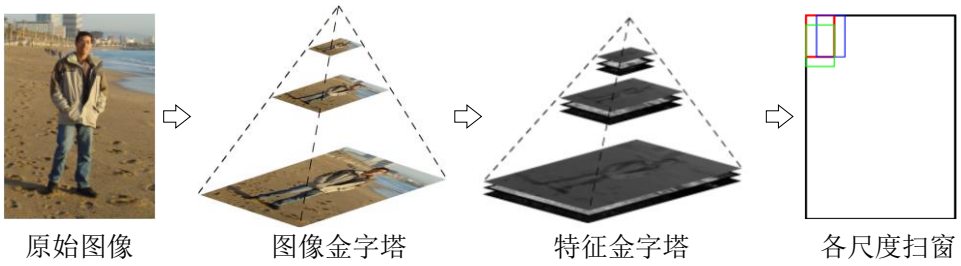


图 2-2 特征金字塔的构建和扫窗

上逐像素扫描窗口并从其在对应的位置提取特征。建立多尺度金字塔可以解决目标物体大小不一致的问题；逐像素扫描则穷举出了每个可能位置而无遗漏。但遗憾的是扫窗会产生大量的窗口，对于一幅标清的720×480图像，构建四阶金字塔，每阶八层，会产生百万级别的窗口。扫窗产生的窗口如此之多严重影响了分类速度，不利于更复杂表示方式的应用。另一方面，已经很复杂的特征学习并不适合于构建特征金字塔。因此近几年来一些待检测目标候选区域的方法涌现出来[30]。通过候选区域可以将候选窗口的个数降低到数千的级别，然后对每个窗口单独提取特征并进行分类以确定待检测目标的位置。候选区域一般采用基于分割、置信度或者学习的方式获得。

分割作为数字图像处理的一个基本技术，已有了相对可靠的算法支持。Uijlings[31,32]等人按照一定的准则将分割获得的小像素区域进行分层组合，提出了 Selective Search 提取候选区域。该方法可以将每幅图像的待检测窗口个数降低到两千左右并在[9]中获得了成功应用。遗憾的是这些基于分割的方法都不适合处理颜色区域比较杂乱的目标，例如自行车，而且其提出的候选区域冗余度很高。

置信度方法因使用低维度的特征和构建粗糙的扫窗方式而处理速度极快，因此也成为候选区域提取的重要方式之一。对于一些特定的目标物体，级联检测器中的第一级也具有提取候选区域的作用。在处理所有目标的定位时，Cheng[33]等人从认知的角度提出人眼在辨别物体时是先进行粗定位再进行精检测，因此其将所有目标都缩放到很小的尺度训练一个一般性的弱分类器，构建不等比缩放的图像金字塔以解决目标物体长宽比不一致的问题。置信度候选区域提取的方法中另一个代表性的工作是 EdgeBoxes[34]。作者认为候选窗口一定要满足其内部有很多

完整轮廓，穿越窗口的轮廓应该大部分位于窗口内部。基于这种假设，构建多尺度、多长宽比的扫窗后将置信度大的窗口作为候选窗口。

Erhan[35]等人提出 DeepMultiBox 方法，通过训练图像学习一个 DNN(Deep Neural Network)网络预测候选窗口位置及其置信度。该网络训练时使图像上置信度最高的窗口与标注值匹配最好。虽然实验表明检测性能的提升和DeepMultiBox无直接关系，但是 DeepMultiBox 可以将候选窗口区域降低到数十的量级上。

无论采用何种候选区域提取方式，其目的都是在保证召回率的前提下尽可能地增加候选区域的定位精度、降低候选区域的个数。在全监督检测中，可以降低提取特征的次数，降低检测所需要消耗的时间。在自学习检测器中，能够降低样本的搜索空间。

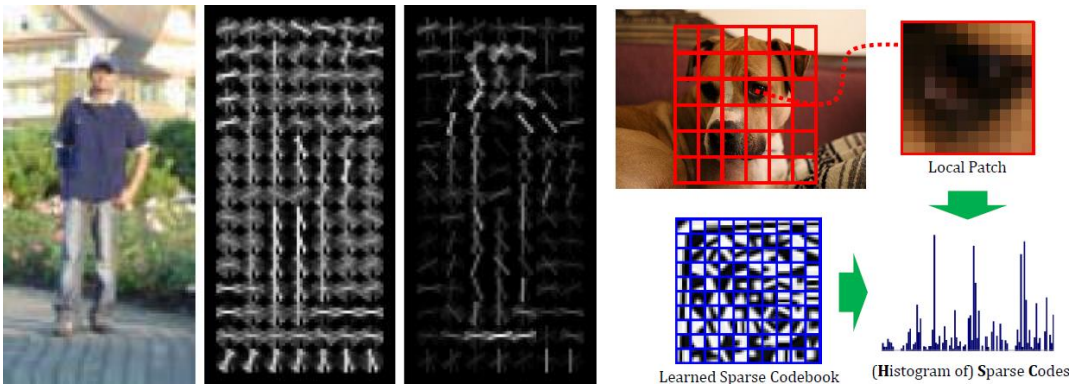
2.3 行人检测的主要方法与分类

2.3.1 全监督行人检测

在很长一段时间里，行人检测都更加关注于特征设计与提取，候选区域都是通过扫窗获得。在这一阶段，行人检测可以分为基于样本特征(feature-based)、基于样本形状(part-based)和基于样本相关线索(multi-cue based)的三大类方法，其中具有代表性的工作分别如图 2-3 到图 2-5 所示。计算简单的特征提取方法对目标的表示能力往往不足，而表示能力好的方法往往计算复杂度很高。在希望不断提高检测性能的背景下，深度学习特征尽管复杂，但依然成为了目标表示的一个趋势。

- 基于样本特征的方法

基于样本特征的方法主要通过设计特征来配合经典的分类器算法对目标进行描述。样本特征又分为手工设计特征和学习特征两种。手工设计特征因其设计简单、检测速度快备受关注。Papagorigiou 和 Poggio[36]提出了使用 Haar 小波函数在样本中提取灰度差特征。Viola 和 Jones[1]在积分图上计算 Haar 特征，并用 AdaBoost 进行分类，达到了实时检测的目的。Dalal[2]等人借鉴图像配准中成功应用的 SIFT (Scale Invariant Feature Transforms) [37]，提出了 HOG (Histogram of Orientated Gradient)局部特征描述子，并用支持向量机(SVM)作为分类器，在行

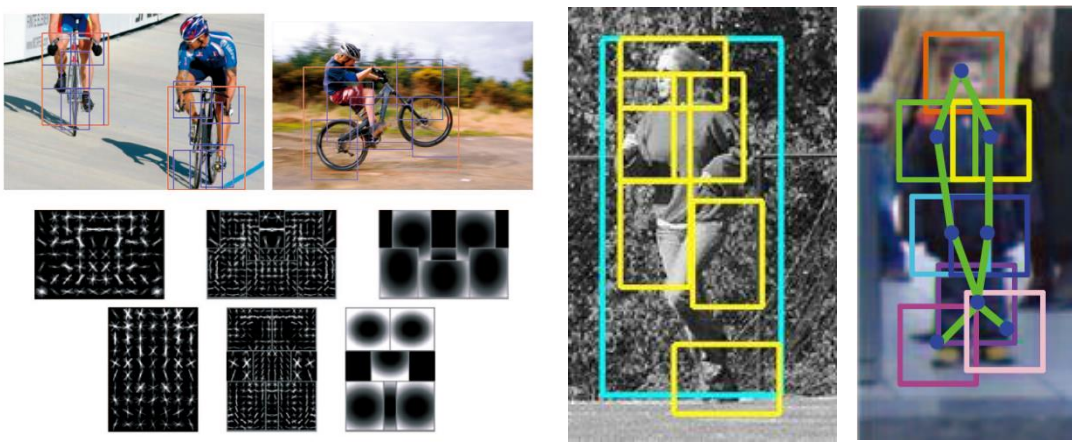


(a) HOG[2] (b) HSC[48]

图 2-3 基于样本特征的典型方法

人检测上获得了前所未有的成功。在 HOG 的基础上一大批优秀的研究接踵而来[38-40]。LBP(Local Binary Pattern)是为描述纹理而设计的特征，并且在纹理检测和人脸识别上取得了很好的效果。Mu[41]等人对 LBP 进行了改进，将其用来描述人体。为了解决行人检测中的遮挡问题，Wang[42] 等人用 HOG 和 LBP 进行特征融合。Dollar[9,43,44]基于前人的研究，结合积分图和 HOG 提出了 ACF，利用级联决策树构建 AdaBoost 分类器，在提高了行人检测精度的同时，在标清图像上检测速度达到 30fps。Nam[45]等发现级联决策树与去相关的特征配合更密切，对 ACF 每个通道进行局部去相关，使得 ACF 在行人检测性能上有很大提高。除此之外，Tuzel[46]用局部块的协方差矩阵对目标进行描述，并将其在黎曼流形上进行分类。

随着稀疏表示和深度学习的发展，通过学习的方式获得一些固定的卷积模板也越来越受研究者们青睐。由于卷积模板无需更新，基于固定卷积模板的特征提取方式仍然属于手工设计特征。Sermanet[47]等人通过字典学习获得卷积模板，在构建双层卷积网络的同时提取全局和局部特征。Ren[48]等人提出通过字典学习对局部块的重构系数进行直方图统计构建 HSC(Histogram of Sparse Coding)，完全采用和 HOG 一样的检测框架，在行人检测上性能比 HOG 提升很多。虽然如此，由于其学习的字典是过完备的，每个局部快的直方图维度是 HOG 的十几倍甚至几十倍。Lim[49]等人用非监督学习的方法获得精确描述目标物体轮廓的中层特征，该方法在训练时依赖于人工标定的轮廓。Zhang[50]等人通过在平均梯度人体上提取 Informed Haar-like 模板对 ACF 进行卷积，和 ACF 一样采用级联决策树构建的 AdaBoost 分类器第一次将 Caltech 行人数据集上高于 50 像素的行人检测错检



(a) 多目标中 DPM[23]　　(b) 行人 DPM[23]　　(c)拓扑结构[51]

图 2-4 基于样本形状的典型方法

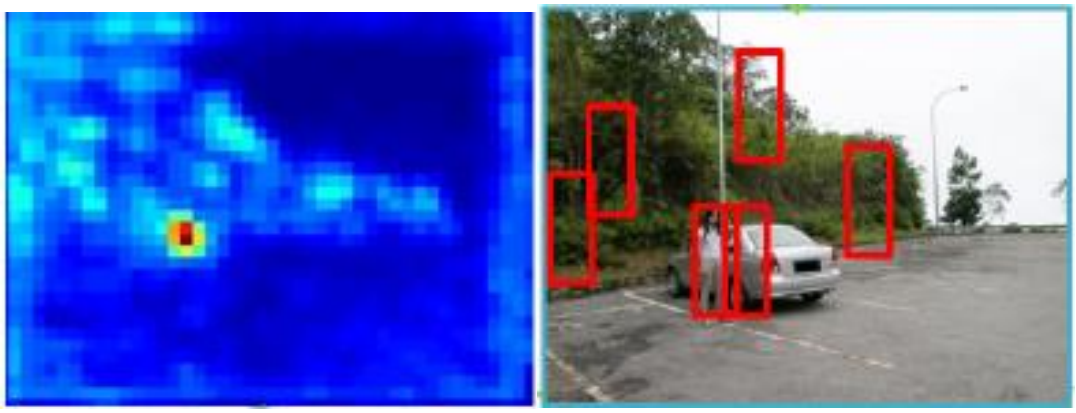
率降到 35%以下。

- 基于样本形状的方法

基于样本形状的方法考虑到目标物体的拓扑结构，将目标物体分为不同的部分进行组合。Felzenszwalb[23]等人提出了形变模型 DPM(Deformable Part-based 方法特别适合处理非刚性目标。两年后，Felzenszwalb[52]等人进一步在 DPM 的基础上加入混合模型(Mixture Model)来处理多视角问题，在 PASCAL 数据集上取得了当时最好的的性能。Sabzmeydani[53]等人将人体分为头、身体和腿三个部分并且每部分都学习一个 Shapelets 特征。Wu 和 Nevatia[54]将人体分成不同的部分，每部分通过 Edgelets 特征进行学习，该方法在目标检测和目标跟踪中都取得了不错的效果。Gao[51]等人通过构建拓扑结构来描述每个局部之间的空间关系，在行人检测上取得了比 DPM 更好的检测率。

- 基于样本相关线索的方法

基于样本相关线索的方法主要采用一些与待检测目标相关的其它信息对目标表示进行辅助。最典型的线索是深度，而深度的获取可以通过立体视觉或者现有的深度获取设备。通过立体视觉获得深度主要应用于交通场景的行人检测。Zhao[55]等人利用基于立体视觉的深度信息将行人从背景中提取出来并实现了行人检测算法。Keller[56]则对深度和灰度图像上分别提取 HOG 特征来提升行人检测性能。Nishiyama[57]通过立体视觉获得感兴趣区域，再采用级联分类器进行目



(a) HOF[58] (b) 上下文关系[59]

图 2-5 基于样本相关线索的典型方法

标检测。利用商业深度获取设备，例如 Kinect，获取深度在辅助行人检测的同时，也可以用于室内多目标检测。类似于[56]，Spinello[60]按照 HOG 的提取方式在 Kinect 获得的深度图上提取 HOD(Histogram of Depth)，并将其与 HOG 融合以提升对行人的描述能力。Bo[61]等人设计核描述子度量三维目标的形状、颜色、梯度相似性并对室内三维目标进行识别。第二种常用来辅助目标检测的线索是运动信息。Dalal[58]继 HOG 之后提出了 HOF(Histogram of Flow)，并验证了其和 HOG 的互补性。Walk[62]等人提出了底层特征的自相似特征 CSS，加入运动信息后表示能力显著提高。Park[63]则利用光流信息将当前帧映射到参考帧上并计算其梯度做为特征描述子。各种各样的上下文信息也被用到辅助检测当中。Dalal[2]在提取 HOG 特征时对样本进行加边，加入了一些背景信息。中科院自动化研究所的 Yan[59]等人将交通场景中人与车之间的相对位置关系加入到他们的多尺度行人检测框架中。Ding[64,65]等人在其方法中加入图像中目标响应集中的信息迭代地训练出一个更好的分类器。

- 基于深度学习的方法

深度学习的兴起促使了 DPM 的进一步发展，Girshick[66]将 DPM 框架中的特征提取、形变模型和混合模型均转换成 CNN 结构。Zhang[29]等人在 R-CNN 结构上，将锚点的长宽比固定为 2 并增加更多多尺度，以适应行人的特点。在训练得到更强表示能力的特征之后，仍然采用 AdaBoost 作为分类器，达到了极佳的检测性能。Cai[67]等人则利用卷积神经网络的多个阶段同时获得候选区域并进行特

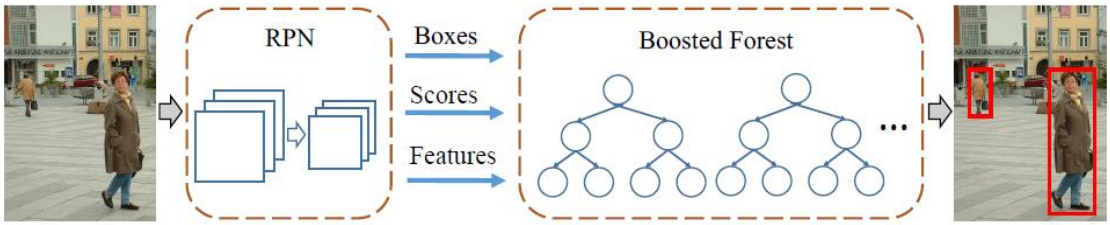


图 2-6 基于深度学习的典型方法[29]

征描述和分类，进行端到端的训练，同一框架下实现了车、行人和骑自行车的人的检测。

2.3.2 自学习行人检测

全监督的行人检测器依赖于完全的样本标注，当在跨数据集时检测性能受到一定的限制，而且随着数据量的快速增长，完全标注样本的代价也越来越高。因此最近几年，使用尽可能少的行人样本训练分类器进行行人检测也越来越受到研究者的关注。迁移学习、在线学习、弱监督方式和无监督方式都是解决样本问题的典型方法。

迁移学习使用目标域的上下文信息以及样本分布来增强源域中预训练的行人检测器。上下文信息[68,69]、置信度传播[69,70]以及虚拟样本[71]可以使迁移学习更平滑。高斯回归[72]和超像素聚类[11]可以用来选择更加可靠的正例行人样本。最大间隔嵌入[73]可以扩大检测器的适应范围。有了迁移学习作为基础，自适应相机[74]的探索也推进了一步。迁移学习虽然不需要在目标域进行样本标定，减少了人工代价，但是它的有效性受到源域和目标域样本差异的影响，这些差异包括行人外观、拍摄视角及光照变化等。当差异特别明显时，迁移学习的有效性会受到极大的挑战，使迁移不平滑或者根本不能够进行迁移。如果能够在同一特定场景中训练自学习行人检测器，就能够有效地避免这种样本差异带来的问题。

在线学习得益于可以使用连续的视频流数据。经典的 detection-by-tracking (DBT)[75]使用离线行人检测器初始化每帧样本位置，再使用时间域的信息降低检测错误。而 Tracking-Learning-Detection (TLD)[76]使用单个准确的样本进行初始化，进行在线跟踪获取更多的样本不断更新检测器以提升检测性能。尽管 DBT 和

TLD 方法很普及，但无论是检测还是跟踪模块的误差都会在整个自学习系统中被放大。

弱监督学习(Weekly Supervised Learning, WSL)中只需要给定图像级的标注，并不需要标定目标的具体位置和大小，也可以一定程度地降低人力标注成本[77]。弱监督学习中假设同一类目标会在某个特征空间存在聚类效应，而背景则很分散。基于这一假设，可以使用聚类信息、跟踪信息、模板匹配、图传播以及多示例学习发现正例行人样本，抑制背景并学习行人检测器。弱监督学习类似于期望最大化，需要迭代地更新样本和检测器。由于缺少标注信息，行人的位置成为一个隐变量，导致弱监督的优化是非凸的，很容易陷入局部极小值而影响检测率。

对于无监督学习，Wu 等人[78]使用在线的局部区域检测器来优化目标检测器，但需要离线地学习一个通用检测器。最近，无监督学习被分解为目标发现和目标跟踪两个互补的单元。Xiao 等人[79]提出一种完全无监督的提取目标候选区域的方法，首先通过聚类并选择易聚成团的样本，然后更新其外观模型，通过初始检测器和时间一致性挖掘更多正例样本。这种无监督的方法可以迭代地产生越来越多的行人候选区域，但不能精确的判定哪个候选区域存在行人。

2.4 本章小结

本章围绕行人检测中的监督信息、框架选择与特征设计三方面的问题，概述了目标检测框架的更迭以及行人检测的主要方法。目标检测框架更迭关注对候选区域进行精细分类这一新框架中候选区域提取的相关工作。本文第三章阐述了使用侧输出残差网络获得更优的行人底层对称信息和边缘特征，第四章使用这些特征设计置信度对冗余候选区域重排序以达到使用更少的窗口数获得更高的召回率定位精度。行人检测方法关注于全监督行人检测和自学习行人检测的发展。本文第五章采用新框架，更着重关注于特征设计，提出了 PCA 卷积特征。得益于候选区域提取，第六章通过渐进优化模型在未标注视频上训练行人检测器，达到了自学习的目的。

第 3 章 基于侧输出残差网络的行人底层特征提取

本章主要关注行人的边缘与对称信息提取。边缘信息可以刻画行人的轮廓，对称信息可以刻画行人的结构，这两个信息都可以被用来描述一个窗口中是否包含行人目标[19]。

基于学习的方法中，边缘和对称性信息提取被当做像素级二分类问题[20,21,80-83]。通过设计合理的像素级特征，可以使用各种学习算法训练分类器对图像中的每个像素进行类别预测，判定其是否位于边缘或者对称轴上。深度学习中，端到端的全卷积神经网络(Fully Convolutional Network, FNC)的提出[84]，使高效的像素级分类成为可能。本章首先简单介绍全卷积网络，然后介绍一种新设计的全卷积网络，侧输出残差网络[20]，对图像进行对称信息提取，在此基础上设计了多分支结构可以同时提取图像的边缘和对称信息[21]，最后在公开数据集上评测侧输出残差网络的性能并做分析。

3.1 全卷积网络简介

全卷积网络最初是为了做图像的语义分割(Semantic Segmentation)，其结构如图 3-1 所示。全卷积网络往往由特征提取、像素分类和上采样三部分构成。在图 3-1 中，对于一幅输入图像，特征提取部分采用 7 层卷积神经网络获得 4096 层的卷积响应图；像素分类仍然利用卷积层学习一个分类器获得 21 层分类结果(PASCAL VOC[3]数据中含有 20 类目标和一类背景)；上采样部分将预测图重新放大到输入图像大小，并采用 softmax 获得每个像素最终的类标。

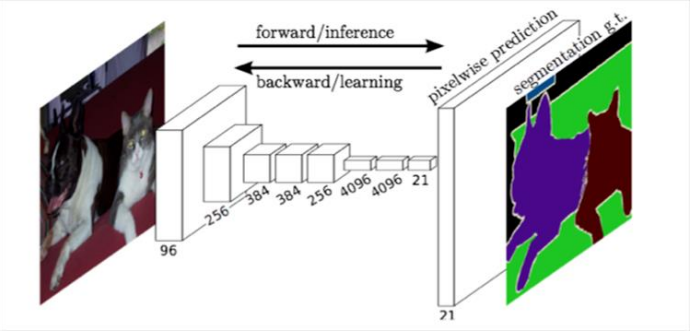


图 3-1 全卷积网络[84]

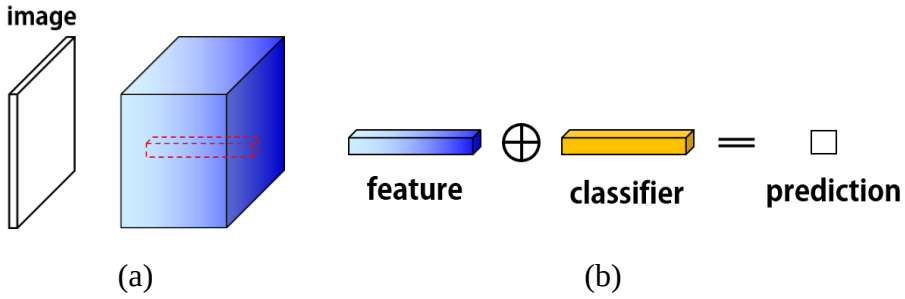


图 3-2 全卷积网络中的像素分类

全卷积网络中的像素分类如图 3-2 所示。不同于卷积神经网络中的全连接层用于图像特征提取[5,35,85-87]，全卷积网络中直接采用卷积响应图作为像素特征。如图 3-2(a)所示，红色长方体为一个像素的特征，并将其表示为 $\mathrm{x}_{l,i,j}=(x_{l,i,j}^{1},x_{l,i,j}^{2},\cdots,x_{l,i,j}^{K})$，其中 l 表示当前卷积所在的层数，i,j 表示像素坐标，$k=1,2,\cdots,K$ 表示响应图的层数。特征层后，可以使用一个 1×1 的卷积作为分类函数 f_l，其维度为 $1\times1\times K$。最终像素的分类结果为：

$$y_{l,i,j}=f\bullet\mathrm{x}_{,l} \tag{3-1}$$

由于卷积神经网络中池化层(Pooling)的存在，最终的分类结果的分辨率小于输入图像的分辨率，所以在最后阶段需要使用上采样使其和输入图像大小一致。在 FCN[84]中最深层的预测结果为输入图像的 $1/32$，有很大的信息损失。多次合并能够在减小这个损失。如图 3-3 所示，先训练 $32\times$ 上采样网络；损失降低到一定程度之后，固定 conv5 的参数，训练 $16\times$ 上采样网络；再固定 conv4 的参数，训练 $8\times$ 上采样的网络；最终以 $8\times$ 上采样的结果作为输出。

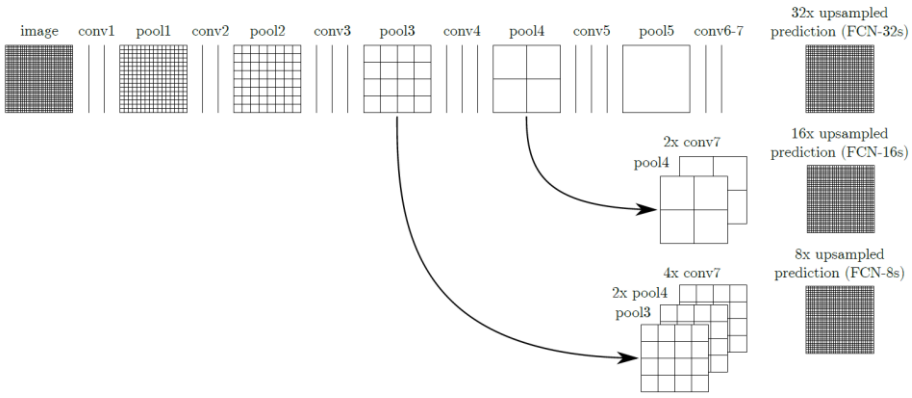


图 3-3 全卷积网络中的多次合并[84]

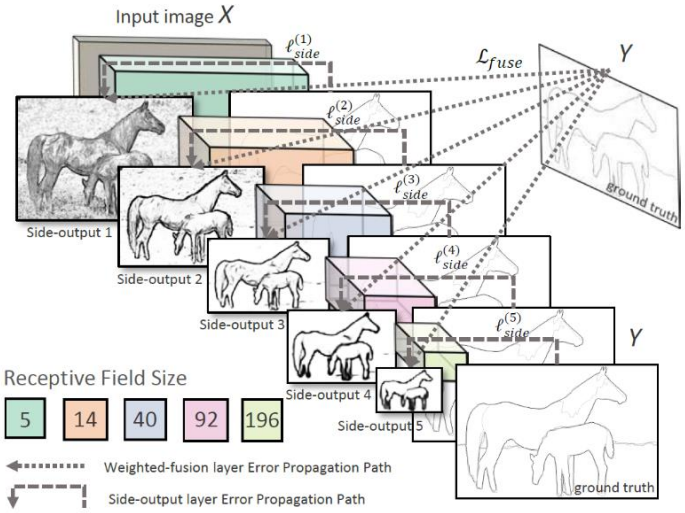


图 3-4 整体嵌套的边缘检测网络 HED 结构[82]

随着深度监督网络(Deeply Supervised Network)的出现[88]，一种端到端的全卷积网络方式，整体嵌套边缘检测网络(Holistically-Nested Edge Detecion, HED),被用在边缘检测问题上，并取得了非常好的效果[82]，如图 3-4 所示。HED 也是用卷积层作为分类器。但不同于 FCN 中的多次合并，HED 使用不同阶段的卷积层作为像素级特征获得侧输出(Side-outputs)，然后再同时监督各个阶段的侧输出，最终将不同侧输出加权平均作为最终的边缘检测结果。假设侧输出为s_i，最终的输出可表示为：

$$o = \sum_i w_i s_i . \tag{3-2}$$

由于卷积神经网络具有先天的多尺度特性，即 5 个卷积阶段的接收野分别为 4、14、40、92 和 196，使 HED 能够同时兼顾卷积神经网络浅层的局部信息和深层的全局信息。在 HED 的基础上，[89]同时使用了多层卷积特征做特征融合，达到了更好的检测性能。

如上所述，不同阶段的卷积层接收野不同，包含不同尺度的信息，使侧输出可以捕捉不同信息。但是也正是因为接收野不同，直接通过固定的加权权重进行合并，不具有尺度的选择效应。也就是说无论真实的尺度对应于哪个接收野，固定的权重都会凸显其固定的接收野所产生的响应。因此，如何设计一个网络结构，能够在一定程度上自适应地选择尺度，将会有效地改进 HED 的性能。侧输出残差网络(Side-output Residual Network, SRN)正是朝着自适应尺度选择而努力[20]。

3.2 基于侧输出残差网络的对称信息提取

本节首先介绍对称性检测以及其发展，然后详细介绍用于对称性信息提取的侧输出残差网络的结构和学习过程。

3.2.1 对称性检测

对称性信息广泛存在于各种视觉目标中，例如天然目标的树木、鸟等和人造目标飞机、管道等。对称的区域和其互相之间的连接关系提供了很强的结构信息。这些信息可以被用来做局部区域分解[90,91]、图像前景分割和提取[92,93]、目标候选区域提取[94]以及文字检测[95]等问题。

早期的对称性检测，也被称为骨架提取，处理的对象是二值图像[96,97]。针对二值图像，可以直接采用图像形态学方法或者细化方法进行骨架提取。最近十年来，有学者选择了一些具有良好对称结构的图像在计算机视觉相关会议上组织竞赛，促进对称性检测在自然场景图像上的发展[22,98]。这个阶段的对称性信息提取往往是采用关键点匹配的方式获取关键点连线的中间点。随着数据集的扩大[20,99-101]，对称性检测也从非监督的提取发展到采用监督方式的学习对称信息检测器。现有的一些对称性数据集如图 3-5 所示。

在自然场景下，无法像二值图像那样对对称轴进行一个很好的数学描述。但是对称轴上的点仍然需要满足：1）左右两侧纹理相近；2）左右两侧到达目标边缘距离相似。通过这样的约束，对称信息仍然能够比较有效地反映出目标的结构。对于行人这类特殊的目标，对称性信息可以描述其形态。

图 3-5 自然场景下对称性检测数据集

3.2.2 侧输出残差网络

侧输出残差网络是通过依次堆叠残差单元(Residual Unit, RU)来实现尺度自适应的。残差单元用来拟合残差单元的输入与真实值之间的差值。当给定一个初始预测值之后，通过残差单元依次拟合预测值与真实值之间的误差，来自适应地选择尺度使得残差最小化。残差单元不仅能够有效地提高 HED[82]的性能，而且也能够使网络收敛得更快。

残差单元：

残差在训练特征提取的网络[86]中获得了极大的成功。不同于用残差来学习更好的图像特征描述，本文的残差单元针对的是预测结果的拟合。残差单元如图 3-6 所示，假设给定一个预测结果 r，映射 $\mathcal{F}(y)$，那么合并后的输出为 $r+\mathcal{F}(y)$。当对残差单元的输入和输出同时加监督时，可以表示为：

$$\begin{cases} r \approx y \\ r+\mathcal{F}(y) \approx y \end{cases}, \tag{3-3}$$

通过公式可以看出，深度监督迫使残差单元的输入和输出都拟合真实值，那么映射 $\mathcal{F}(y)$ 就是在拟合残差单元输入 r 和真实值 y 之间的残差。当将多个残差单元堆叠在一起的时候，其映射 $\mathcal{F}(y)$ 会越来越逼近于 0。在侧输出残差网络当中，映射 $\mathcal{F}(y)$ 和尺度相关，因此残差单元就有了一定的尺度选择能力。在极限情况下，如果残差单元的输入 r 已经最优，那么整个网络的训练会促使 $\mathcal{F}(y)$ 趋于 0。这种通过一定操作使得映射趋于零的方式被证明比用映射拟合真实输出的方式更加容易[86]。

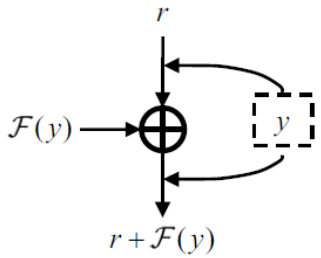


图 3-6 残差单元示意图

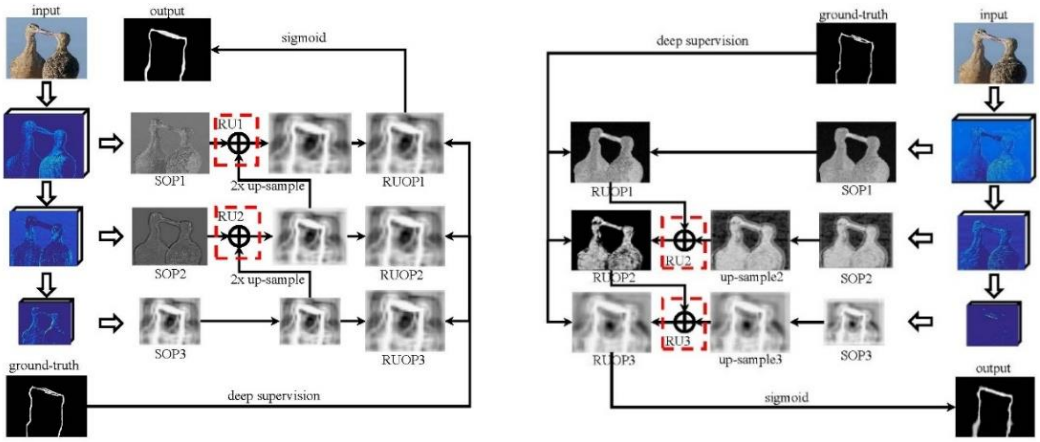


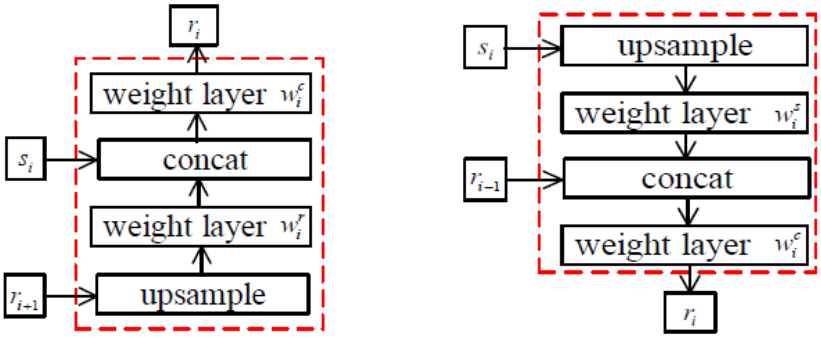
图 3-7 侧输出残差网络结构

网络结构：

根据残差单元堆叠方向的不同，可以构建两种侧输出残差网络结构，分别为自深到浅的结构和自浅到深的结构，如图 3-7 所示。本文选用 VGG[102]作为根网络提取特征。VGG 是由五个阶段构成的网络，每个阶段包含二到三层卷积层，每个阶段之间有池化层。为了方便表示，图 3-7 中只画了三个阶段。在图中，每个残差单元的标号和其对应的侧输出(SOP, side-output)的标号一致，并且第i个残差单元输出被记做 RUOP_i (RU output)。

*自深到浅：*自深到浅结构中的残差单元是由卷积神经网络的最深层到最浅层依次堆叠，如图 3-7(a)所示。分别将第 i 个侧输出表示为 s_i，被堆叠的第 i 个残差单元的输入和输出表示为 r_{i+1} 和 r_i。对于第一个残差单元 RU2，其输入直接被设置为最深层的侧输出，即 $r_3 = s_3$；对于其它残差单元，输入为其上一个残差单元的输出。将图 3-7(a)中的结构带入公式(3-3)，除了最深层的侧输出以外，其它侧输出均在依次拟合残差。最终，只需要将最浅层的残差单元的输出作为最终的输出。对于边缘或者对称性这种二分类问题，本文还使用 sigmoid 函数将输出的结构归一化到 0 到 1 之间。

在自深到浅的结构中，残差单元的实现如图 3-8(a)所示。在这个结构中，残差单元的输出图像 r_i 的大小和侧输出图像 s_i 大小一致，需要将其输入图像 r_{i+1} 进行



(a) 自深到浅　　(b) 自浅到深

图 3-8 残差单元的实现

2×的上采样。本文没有采用直接叠加 s_i 和上采样后的 r_{i+1} 的方式，而是用了一个 1×1 的卷积层对其进行加权平均。通过图 3-8(a)，残差单元可以表示为：

$$r_i = w_i^c(s_i + w_i^r r_{i+1}), \tag{3-4}$$

其中 w_i^c, w_i^r 是卷积层的参数。将公式(3-4)带入公式(3-3)可以得到残差映射 $\mathcal{F}_i(y)$：

$$\mathcal{F}_i(y) = w_i^c \cdot s_i + (w_i^r w_i^c - 1) r_{i+1}. \tag{3-5}$$

在卷积神经网络当中，更深的层由于具有更大的接收野，所以其采集的信息是全局的高层次的信息。最深层的侧输出也就和真实值更加接近，也就是说采用最深层相比于最浅层而言是一个更好的初始值。

*自浅到深：*自浅到深的结构如图 3-7(b)所示，其残差单元的实现如图 3-8(b)所示。在这个结构中所有的侧输出都需要被上采样的输入图像大小，再利用残差单元进行合并。同样的，可以得到残差映射 $\mathcal{F}_i(y)$：

$$\mathcal{F}_i(y) = w_i^s w_i^c \cdot s_i + (w_i^c - 1) r_{i+1}. \tag{3-6}$$

学习过程

给定对称性检测的训练数据集 $S = \{(X_n, Y_n)\}_{n=1}^N$，包含 N 对训练样本。$X_n = \{x_j^{(n)}, j = 1, \cdots, T\}$ 和 $Y_n = \{y_j^{(n)}, j = 1, \cdots, T\}$ 分别表示一幅输入图像和其对称性的真实值，每幅图像含有 T 个像素。$y_j^{(n)} = 1$ 表示第 j 个像素位于对称轴上，反之 $y_j^{(n)} = 0$ 表示该像素为背景像素。假设特征提取网络的模型参数为 $\mathbf{W}$ 并且含有 M 个侧输出，那么将会有 $M-1$ 个残差单元被堆叠在一起。本文中采用 VGG 做为特征提取

网络，共有$M=5$个侧输出。学习过程的建模以图 3-7(a)中的自深到浅为例，自浅到深有类似的过程。

对于作为初始值的侧输出，对应于图 3-7(a)中 RUOP3，其损失为：

$$\mathcal{L}_b(\mathbf{W},w^b)=-\beta\sum_{j\in Y_+}\log\Pr(y_j=1\mid X;\mathbf{W},w^b)-(1-\beta)\sum_{j\in Y_-}\log\Pr(y_j=0\mid X;\mathbf{W},w^b), \tag{3-7}$$

其中w^b为其分类器参数，损失的权重为$\beta=|Y_+|/|Y|$，$|Y_+|$和$|Y_-|$分别表示正例像素和反例像素的个数。使用 sigmoid 将分类结果概率化后得到$\Pr(y_j=1\mid X;\mathbf{W},w^b)\in[0,1]$。这个值用来衡量一个点有多大的可能性位于对称轴上。对于第i个残差单元，其损失为：

$$\mathcal{L}_i(\mathbf{W},\theta^i,w^i)=-\beta\sum_{j\in Y_+}\log\Pr(y_j=1\mid X;\mathbf{W},\theta^i,w^i)-(1-\beta)\sum_{j\in Y_-}\log\Pr(y_j=0\mid X;\mathbf{W},\theta^i,w^i) \tag{3-7}$$

其中$\theta_i=(w_i^c,w_i^s)$是残差单元中卷积层的参数，w^i是分类器参数。总的损失可以表示为：

$$\mathcal{L}(\mathbf{W},\theta,w)=\alpha_1\mathcal{L}_b(\mathbf{W},w^b)+\sum_{i=2}^{M}\alpha_i\mathcal{L}_b(\mathbf{W},\theta^i,w^i). \tag{3-8}$$

训练阶段可以描述为优化参数使损失最小化：

$$(\mathbf{W},\theta,w)^*=\arg\min\mathcal{L}(\mathbf{W},\theta,w). \tag{3-9}$$

当学习到一个局部最优的模型之后，在测试阶段，给定一幅图像X，其对称性结果为：

$$\hat{Y}=\Pr(y_j=1\mid X;\mathbf{W}^*,\theta^*,w^*). \tag{3-10}$$

一方面，侧输出残差网络通过序列残差拟合来达到降低分类错误的目的。这种思想在自动控制中已经被证实比直接估计误差要有效得多[103]。通过这种序列残差拟合，侧输出残差网络使误差流随着卷积神经网络中的尺度变化依次减小。另一方面，侧输出残差网络结构还是一种特殊的分类器聚合[104]。每个阶段的卷积网络的侧输出可以看做是一个弱分类，最终的输出相当于整合了多个弱分类的结果。

3.3 基于多分支结构的边缘与对称信息提取

除了对称信息之外，边缘信息也是经常被用到的底层特征。和对称性一样，

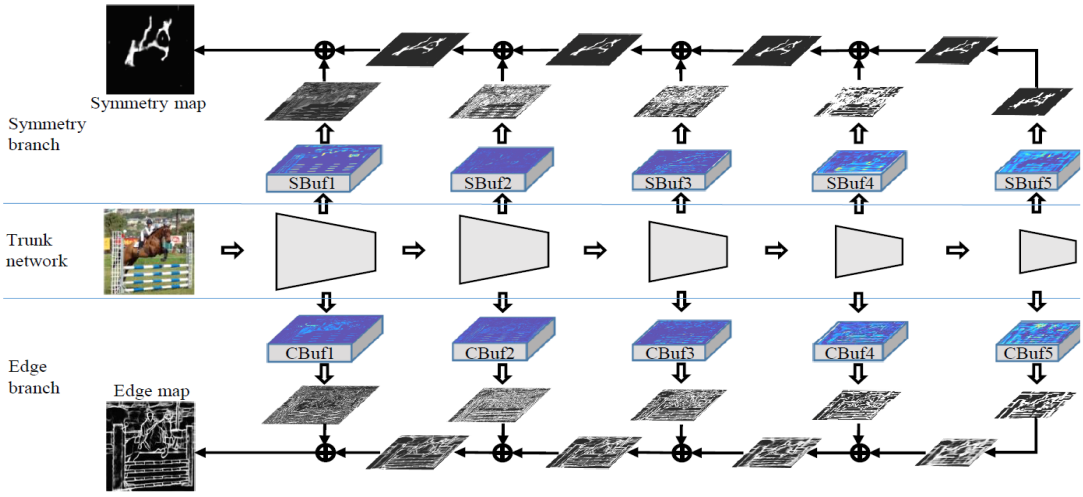


图 3-9 多分支侧输出残差网络结构

边缘检测也可以被看做一个像素级的二分类问题，那么侧输出残差网络也就能够应用到边缘检测这个问题上。

如果单独用侧输出残差网络训练一个边缘检测的模型和一个对称信息提取模型，那么图像在提取像素级特征的时候，计算冗余度很高。本节中，通过设计多分支侧输出残差网络和一定的训练策略，使模型能够同时提取输入图像的边缘信息和对称信息，如图 3-9 所示。侧输出残差网络可以很自然地分为两部分：根网络用来提取像素级特征，分支网络用来分层融合多个侧输出。当分支网络的任务是对称信息提取时，其和根网络构成了一个完整的侧输出残差网络；当分支网络的任务是边缘检测时，其和根网络又构成了一个新的侧输出残差网络。当训练数据没有同时标定的对称性和边缘真实值时，多分支侧输出网络需要分阶段

地进行训练，即依次迭代地训练对称性分支和边缘分支。这种分阶段的训练会使根网络的参数在任务更替时发生很大的变动，因此本文在每个分支网络之前先加一层缓冲卷积层。缓冲卷积层能够使像素级特征提取的根网络的参数变动减小，同时也可以将根网络的像素特征映射到一个新的特征空间当中。缓冲层已经在多尺度 RPN(Region Proposal Network)中得到了很好的使用[67]。

在训练阶段，多分支侧输出残差网络会包含两个损失函数：

$$\mathcal{L}_T = \mathcal{L}(\mathbf{W}_T, \theta_S, w_S) \tag{3-11}$$

$$\mathcal{L}_E = \mathcal{L}(\mathbf{W}_T, \theta_E, w_E) \tag{3-12}$$

其形式如公式(3-8)所示，其中 $\mathbf{w}_T$ 表示根网络的参数，θ_S, θ_E 表示对称性分支和边缘分支的参数，w_S, w_E 表示对称性分支和边缘分支的分类器参数。训练时就依次最小化这两个损失函数，

$$\arg\min \mathcal{L}(\mathbf{W}_T, \theta_S, w_S) \tag{3-13}$$

$$\arg\min \mathcal{L}(\mathbf{W}_T, \theta_E, w_E) \tag{3-14}$$

测试阶段，输入一幅图像，同时在两个分支中进行前向传播，以获得对称性检测结果和边缘检测结果。

多分支侧输出残差网络除了继承了残差网络的优点之外，还有如下优点：1) 底层像素级特征提取共享，在降低了计算冗余度的同时，能够利用跨数据集的训练数据进行训练。根网络用到了更多的训练数据，其特征描述能力得到了加强。2) 虽然每个分支网络增加了缓冲卷积层，但是多次共用根网络，仍然节省了模型的存储空间。3) 在测试阶段可以同时获得对称信息和边缘信息，增加了计算速度。

3.4 实验结果及分析

侧输出残差网络是在 HED[4]的基础上做的改进，在实验分析之前，首先直

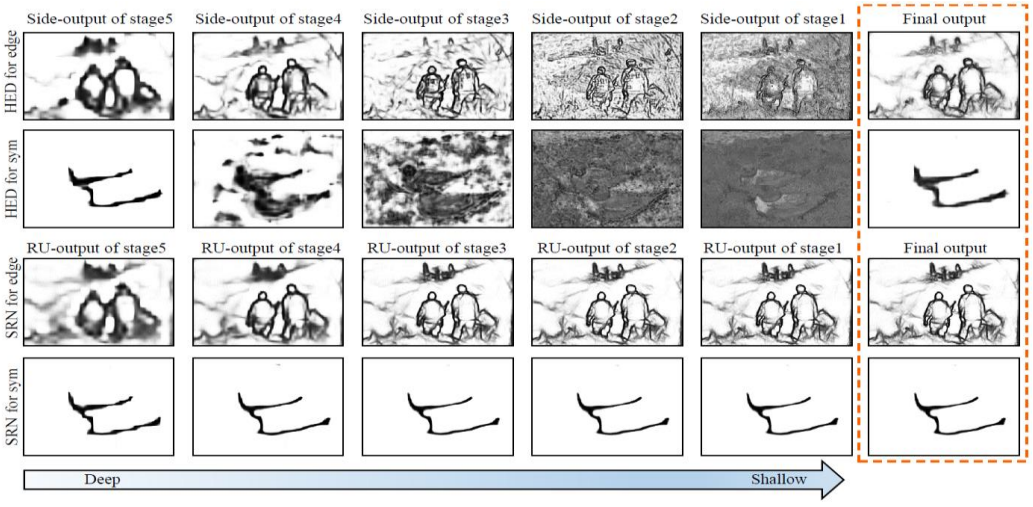


图 3-10 SRN 和 HED 的边缘与对称性检测结果比较

观的对侧输出残差网络和 HED 的结果进行比较，如图 3-10 所示。图中自左向右分别为最深层到最浅层的输出以及最终的输出，上面两行展示了 HED 的结果，下面两行展示了侧输出残差网络的结果。从图中可以直观得看到，侧输出残差网络自最深层到最浅层的输出依次变好，具有一定的尺度选择能力。而 HED 采用的是不同侧输出的加权平均，噪声点高于侧输出残差网络。

3.4.1 实验配置

数据集：实验中评估的数据集包括：对称性检测数据集 SYMMAX[82]，WH-SYMMAX[99]，SK506[100]，以及 Sym-PASCAL[20]；边缘检测数据集 BSDS500[83]。

SYMMAX 由 300 幅自然场景图像构成，其中训练数据 200 幅，测试数据 100 幅，同时标注了前景和背景的对称信息。虽然更多时候，计算机视觉任务更加关注前景目标，但对于底层对称性信息提取，这个数据集具有更大的泛化能力。

WH-SYMMAX 是由 328 幅马匹的样本图像构成，其中 228 幅用作训练数据，100 幅用作测试。这个数据集只关注马这一类目标，而且图像中背景占比不大，马的姿态都为侧面。

SK506 是 WH-SYMMAX 同一个作者提出的多目标数据集，包含 300 幅训练图像和 206 幅测试图像。图像中背景仍然被裁切掉。该数据集包含 16 类目标，其中飞机、行人、鸟和长颈鹿占比比较大。

Sym-PASCAL 数据集包含 648 幅训练图像和 787 幅测试图像。该数据集直接使用的自然场景图像，具有很高的挑战性，包括：目标姿态多样，多目标共现，目标与目标之间、背景与目标之间的遮挡，背景占比大而复杂等。

BSDS500 数据集是很典型的边缘检测数据集，由 200 幅训练图像、200 幅验证图像以及 200 幅测试图像构成。

超参数设置：选择 VGG-16[102]作为像素级特征提取网络并进行再训练(fine-tune)，训练阶段优化的超参数为：每次迭代输入图像数量为 1；对于自然场景图像学习率为10^{-8}，对于样本图像设为10^{-6}；侧输出损失权重为 1；优化动量为 0.9；权重正则项为 0.002；分类器卷积层初始化为 0；总迭代 18000 次。在测试阶段，还需要用非极大值抑制算法[105]作后处理。

评测标准：PR 曲线和 F-measure 被用作评测标准，F-measure 的计算如公式

(3-15)所示：

$$F = \frac{2PR}{P+R}, \tag{3-15}$$

其中 P, R 分别表示精度(Precision)和召回率(Recall)。从 0 到 1 等间隔给定一系列阈值之后，可以计算不同阈值下的精度和召回率，绘制 PR 曲线。PR 曲线中越靠近左上角，F-measure 越大，表示着算法的性能越好。

3.4.2参数选择

侧输出残差网络中的参数选择主要包括三个方面：采用自深到浅还是自浅到深的结构，训练数据扩充采用单尺度还是多尺度，是否融合 conv1 的侧输出的结果。实验结果如表 3-1 所示。通过两两对比发现，自深到浅的结构的性能比自浅到深的结构更好，因为自深到浅的结构初始化得更好；做单尺度的数据扩充要比多尺度好，因为对于对称性和边缘而言都只有一像素宽度，进行放大时会让线宽变大，缩小时可能使一些线不连续；无论是否包含 conv1 的侧输出，性能相似，conv1 的接收野只有 5 个像素，对于 HED 而言确实不足以在这么小的接收野下获得对称信息的响应，但是对于侧输出残差网络而言，由于其由粗到精的结构使 conv1 是否存在对结果基本不产生影响。

在后面的实验当中，选择自深到浅的网络结构，单尺度扩充，不使用 conv1 的侧输出。

Architecture	Augmentation	Conv1	F-measure
Shallow-deep	1x	with	0.381
		w/o	0.397
	0.8x, 1x, 1.2x	with	0.371
		w/o	0.396
Deep-shallow	1x	with	**0.443**
		w/o	**0.443**
	0.8x, 1x, 1.2x	with	0.384
		w/o	0.397

表 3-1 侧输出残差网络的参数选择

3.4.3 对称性检测实验结果及分析

在对称性评测中先在 Sym-PASCAL 数据集上进行详细分析，然后评测其它三个数据集的性能，最后给出收敛性的比较。

Sym-PASCAL 上的实验结果：Sym-PASCAL 数据集上的对称性检测结果如图 3-11 和表 3-2 所示。所有方法的实验结果都是根据原文作者提供的代码重新运行之后获得的。

从表 3-2 可以看到，传统的方法不仅性能不是很好，而且需要大量的计算时间，主要是因为像素级的特征提取本来计算量就很大，再考虑到多尺度和多方向之后会成倍增加特征提取的计算复杂度。传统方法中，性能最高的是多实例学习(Multiple Instance Learning, MIL)[99]，F-measure 达到了 0.174。运行最快的是 Lindeberg[81]，在 CPU 平台上平均需要 5.79 秒处理一幅图像。而 Levinshtein[106]，Lee[107]和 Particle Filter[108]则需要更多的计算时间。

相比于传统方法，端到端的深度学习方法不仅在性能上有很大的提高，而且计算速度非常快。对于基准方法 HED 而言，F-measure 可以达到 0.369，而处理一幅图像在 GPU 上只需要 0.1 秒。FSDS[100]是在 HED 的基础上加入尺度信息以增加分类的准确性，其性能高于 HED 达到了 0.418，计算速度仅仅比 HED 慢了 0.02 秒。本文的侧输出残差网络也是在 HED 的基础上进行改进，但并不需要使用尺度信息，F-measure 达到了 0.443。相比于 HED，提高了 7.4%，即使相比于 FSDS，性能也高出了 2.5%。

端到端的深度学习方法在 Sym-PASCAL 上的对称性检测结果如图 3-12 所示。从最左侧开始，依次展示了简单背景下单目标、复杂背景下单目标、简单背景下双目标、复杂背景下双目标、复杂背景下双目标的遮挡，以及多目标情况下的对称性检测结果。从图中可以看出，侧输出残差网络的检测结果和真实值更加一致。

*分阶段方法：*由于自然场景下对称性检测数据集的标定一般都是依赖于语义分割的标定结果，所以本文还比较了先使用最好的语义分割方法获得分割结果，再在二值分割结果上提取骨架的方式。与此同时，由于 FSDS 方法提出时主要使用样本式数据集，本文还比较了先使用目标检测方法获得检测结果，再使用 FSDS

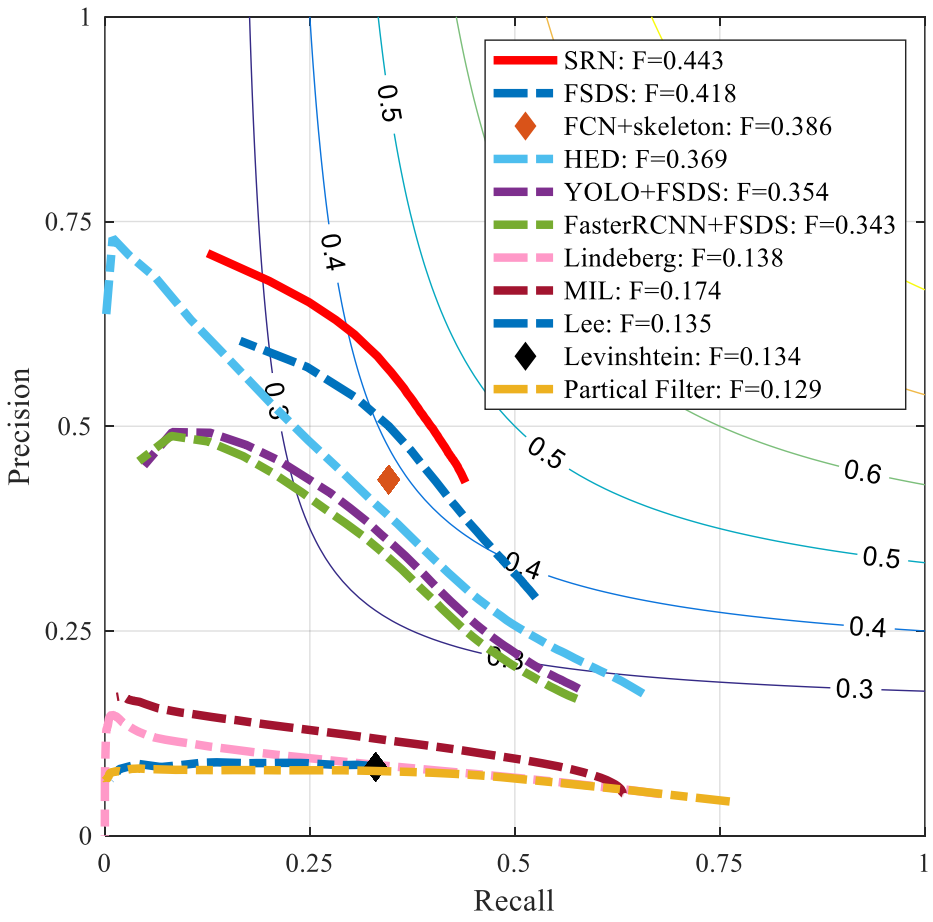


图 3-11 在 Sym-PASCAL 数据集上的 PR 曲线

methods	F-measure	Runtime(s)
Particle Filter[108]	0.129	25.3
Levinshtein[106]	0.134	183.87
Lee[107]	0.135	685.94
Lindeberg[81]	0.138	5.79
MIL[99]	0.174	80.35
HED (baseline)[82]	0.369	**0.10**
FSDS[100]	0.418	0.12
FasterRCNN[109]+FSDS[100]	0.343	0.33
YOLO[26]+FSDS[100]	0.354	0.12
FCN[84]+skeleton[110]	0.386	0.76
SRN (ours)	**0.443**	0.12

表 3-2 在 Sym-PASCAL 上对称性检测性能对比

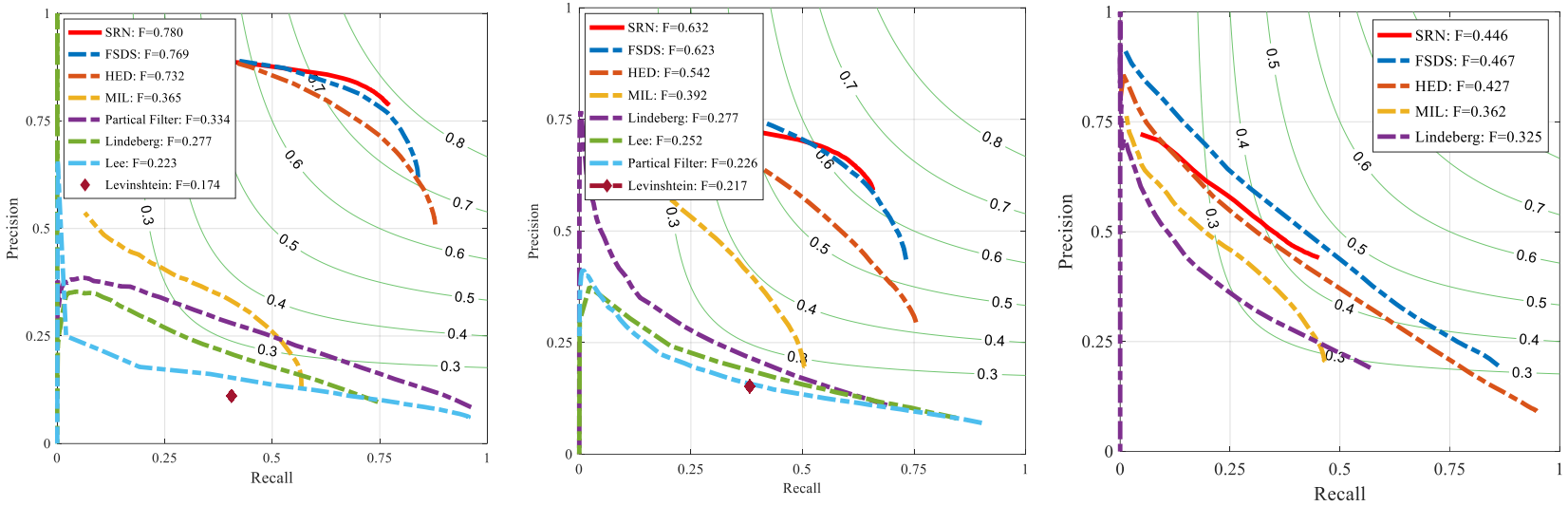
图 3-12 端到端的深度学习方法在 Sym-PASCAL 上的检测结果

进行对称性检测。从图 3-11 和表 3-2 中可以看出，这种分阶段的方式性能并不如端到端的方式，主要原因在与语义分割和目标检测都仍然是未能完全解决的计算机视觉任务，误差会进行累积。同时分阶段的方式由于需要采取顺序两步走策略，相比于端到端的方法，需要更多的计算时间。

其它数据集检测结果：本文也比较了侧输出残差网络以及各种其它方法在另外三个公开数据集上的检测结果，如图 3-13 和表 3-3 所示。和 Sym-PASCAL 数据集中一样，端到端的深度学习方法表现出了很好的性能，在 SK506 和 WH-SYMMAX 上，侧输出残差网络取得了最好的性能。在 SYMMAX 上侧输出残差网络相比于 FSDS，性能很接近，但高于基准方法 HED 的检测性能。

收敛性验证：在最后，本小节展示了侧输出残差网络以及 HED 的损失曲线，如图 3-14 所示。图中同时展示了随着训练迭代次数的增加，损失函数值的变化，以及在此情况下得到的模型的检测结果。

从图中可以看到，HED 的损失曲线一直处于振荡状态，但随着迭代次数的增加，HED 的检测性能趋于平稳。而侧输出残差网络的损失曲线有很明显的收敛趋势，检测性能和 HED 一样上升一段时间之后趋于平稳。一方面侧输出残差网络的损失曲线可以指导训练是否终止，另一方面得益于依次拟合残差，侧输出



(a) WH-SYMMAX (b) SK506 (c) SYMMAX

图 3-13 公开数据集上对称性检测方法的 PR 曲线

datasets	Levinshtein[106]	Lee[107]	Lindeberg[81]	Particle Filter[108]	MIL[99]	HED[82]	FSDS[100]	SRN(ours)
WH-SYMMAX	0.174	0.223	0.277	0.334	0.365	0.732	0.769	**0.780**
SK506	0.217	0.252	0.227	0.226	0.392	0.542	0.623	**0.632**
SYMMAX	--	--	0.360	--	0.362	0.427	**0.467**	0.446

表 3-3 公开数据集上各种对称性检测方法的 F-measure

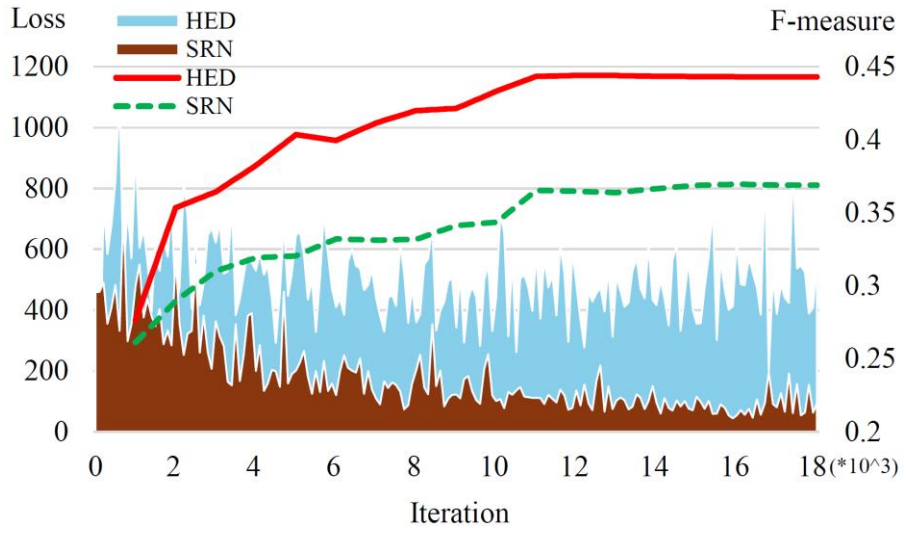


图 3-14 训练阶段损失曲线以及检测阶段 F-measure

残差网络能够较快得达到与 HED 可比的性能。

3.4.4 边缘检测结果及分析

边缘检测可以通过 F-measure 的最大值(ODS, Optimal Dataset Scale)，F-measure 平均值的(OIS, Optimal Image Scale)，以及平均精度(AP, Average Precision)三个指标进行评测。在 BSDS500 数据集上的边缘检测结果如图 3-15 和表 3-4 所示。由于 BSDS500 每幅图像都是由数位标注者标注的，因此其能够得到人为标注的 ODS=0.8，OIS=0.8。这也可以当做各种方法所能达到的性能的上限。而经过研究者们一代代的努力，边缘检测也越来越接近人为标注的性能。Canny 算子边缘检测[111]虽然精度不是很高，但因其效率是一个很大的优点。相比于手工设计特征时代的 SE[105]和 Sketch Token[49]，深度学习的方法获得了更好的性能。非端到端的 DeepContour[112]虽然性能有了提升，但是计算速度很慢，而端到端的 HED 以及侧输出残差网络不仅性能高而且速度计算速度快。HED 的 ODS 达到了 0.780，侧输出残差网络达到了 0.782。正如前面分析，边缘检测性能已趋近人的性能，因此即使很少的性能提升也有一定的意义。

3.4.5 多分支结构结果及分析

在多分支侧输出残差网络中，我们仍然使用 VGG[102]作为基网络。由于边缘响应并不需要很大的感受野，所以我们同时使用了 VGG 的所有五个卷积阶段的侧输出构建侧输出残差网络分支。实验中缓冲卷积层的输出层数和其连接的上一

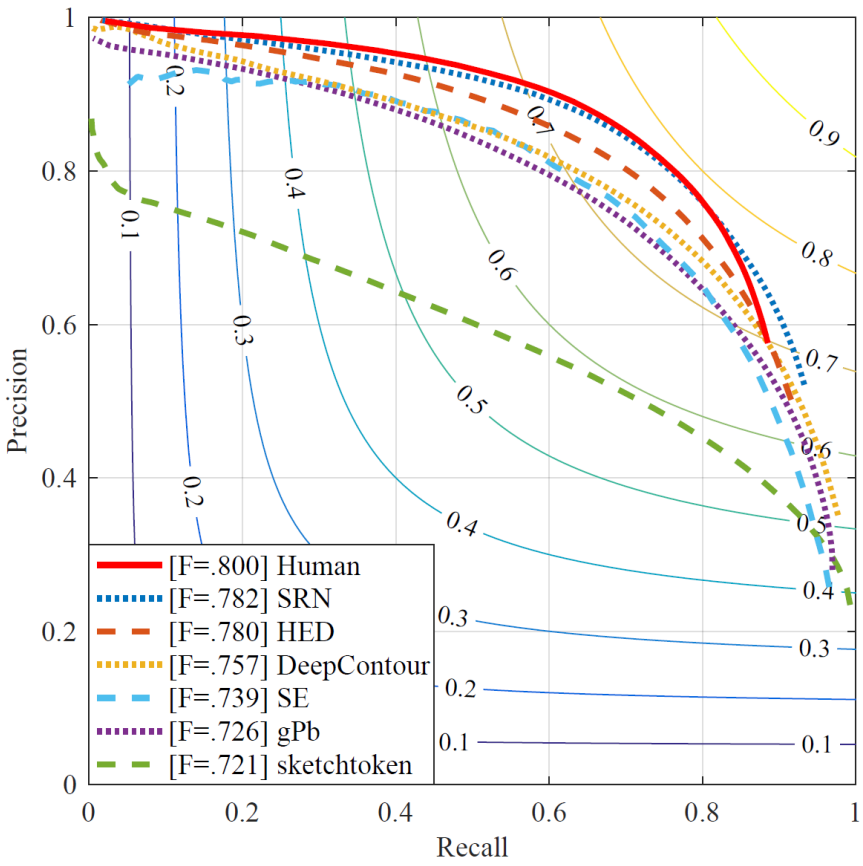


图 3-15 在 BSDS500 数据集上的边缘检测 PR 曲线

Methods	ODS	OIS	AP	FPS
Human	0.800	0.800	--	--
Canny[111]	0.590	0.620	0.578	**15**
Sketch Tokens[49]	0.721	0.739	0.768	1
SE[105]	0.739	0.759	0.792	2.5
DeepContour[112]	0.757	0.776	0.790	0.03
HED[82]	0.780	0.797	**0.814**	2.5
SRN (ours)	**0.782**	**0.800**	0.779	2.3

表 3-4 BSDS500 上的边缘检测结果比较

层卷积层层数一致。训练时共分为五个阶段，每个阶段都训练一次边缘分支和一次对称性分支。五个阶段的迭代次数分别为 8000、6000、4000、2000 和 1000，前两个阶段学习率设为10^{-6}，后三个阶段设为10^{-8}。

本小节首先评测了缓冲层在保护基网络参数大范围浮动上的作用，如图 3-16

所示。我们定义参数浮动的衡量标准为：

$$d = \left\| w_l^{t+1} - w_l^t \right\|^2 \tag{3-16}$$

其中 w_l^t 表示第 t 个训练阶段基网络中第 l 个卷积层的参数，浮动越大，则公式(3-16)的值就越大。从图 3-16 可以看出在单独训练某一个分支时，这个浮动会逐步降低，例如从刚开始迭代到完成 E1 次迭代这一阶段。当从一个分支转换到另一个分支时，参数浮动会变大，例如完成 E1 后进入 S1 阶段。由于缓冲层的存在，可以看到浮动虽然在每次分支转换时会有较小的增长，但整体的趋势越来越趋近于 0。

多分支侧输出残差网络结构的检测性能评测仍然用上文提到的数据集，如表 3-5 所示。在表的上半部分，展示了单独使用侧输出残差网络在某个特定数据集上的结果。下半部分展示了使用边缘检测数据集 BSDS500 分别和 SYMMAX、WH-SYMMAX、SK506 和 Sym-PASCAL 进行组合训练多分支侧输出残差网络的结果。对于(BSDS500, SYMMAX)的组合，边缘检测的性能由 0.782 增长到 0.785，对称性检测性能由 0.446 增长到 0.464。其它三种组合(BSDS500, WH-SYMMAX)、(BSDS500, SK506)和(BSDS500, Sym-PASCAL)上也表现出了类似的性能增长。

两个单独任务的性能并没有因为同时完成而产生不利影响，反而都有所加强。一方面是因为网络在训练过程中，总的训练次数增加了，由单独的侧输出残差网络的 20000 次增加到多分支结构的 42000 次，使得基网络被更好地训练。另一方面是因为多数据集的使用，使基网络的特征表达能力更强。这种训练方式也为数据量小的图像到图像转换的任务提供了可能，可以使用其它任务的数据进行底层像素特征的增强。

3.4.6行人底层信息提取

INRIA 数据集上的部分图像对称性提取结果如图 3-17 所示。可以看到无论是单目标还是多目标，无论大尺度还是小尺度，都能比较好的捕获了行人所在的区域，形成了低层的骨架特征，进而为目标定位提供了更加可靠的线索。

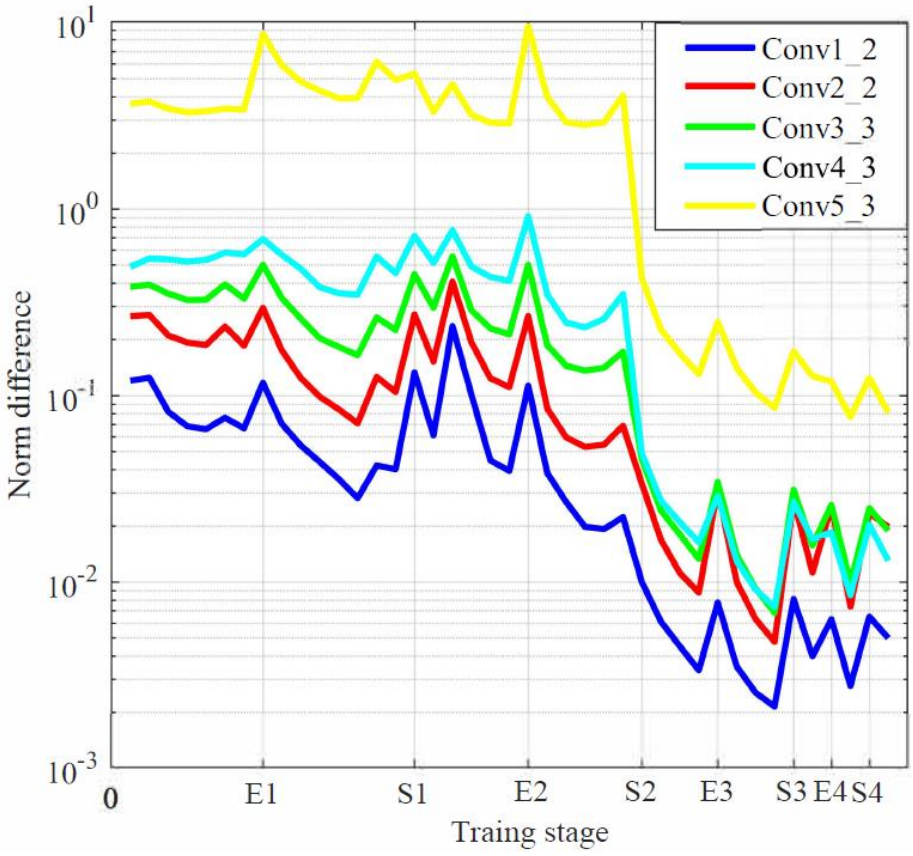


图 3-16 多分支侧输出残差网络中基网络的参数变动

	Datasets	ODS (edge)	F-measure (symmetry)
SRN	BSDS500	0.782	--
	SYMMAX	--	0.446
	WH-SYMMAX	--	0.780
	SK506	--	0.632
	Sym-PASCAL	--	0.443
MTSRN	(BSDS500, SYMMAX)	0785	**0.464**
	(BSDS500, WH-SYMMAX)	0.779	**0.807**
	(BSDS500, SK506)	**0.786**	**0.639**
	(BSDS500, Sym-PASCAL)	0.784	**0.453**

表 3-5 多分支侧输出残差网络在边缘和对称性数据集上的评测结果

图 3-17 在 INRIA 数据集上对称性检测结果

3.5 本章小结

本章提出了侧输出残差网络（SRN）并用于行人目标底层特征提取。侧输出残差网络利用卷积神经网络的多尺度特性，通过残差单元依次堆叠估计初始值真实值之间的残差，自适应地选择合理的尺度不断减小输出结果的损失，提升了对称性和边缘的检测性能。此外通过构建多分支侧输出残差网络结构，可以同时提取输入图像的对称和边缘信息，降低了计算复杂度，减少了模型存储空间。

第 4 章 基于贝叶斯得分重排序的候选区域提取

由于对称性和边缘信息都是图像的底层像素特征，往往在公开数据集上训练的模型就可以直接用于其它数据集的图像，并不需要进行重新训练。因此在快速获得底层的边缘和对称性信息之后，可以用其辅助无监督的候选区域提取。本章首先简单介绍无监督候选区域提取的典型方式，然后通过这些方式的互补性提出了基于贝叶斯得分重排序的候选区域提取方法，最后在公开数据集上进行评测并展示了该方法在行人候选区域提取上的结果。

4.1 无监督候选区域提取的典型方法

无监督候选区域提取的典型方式有两类：基于超像素合并的方式和通过合理的假设给定置信度排序的方式。本小节分别介绍这两类方式中的典型方法 Selective Search[31,32]和 EdgeBoxes[34]。

4.1.1 Selective Search

超像素合并方式产生目标候选区域首先通过快速图割[113]将图像分解成很多超像素区域，然后通过合并超像素获得一些候选区域。这种搜索是 NP 难问题。Selective Search 一方面通过设计一些合理超像素对的相似性度量，贪心地进行合并，增加搜索速度。另一方面通过采用不同的颜色空间产生不同的超像素图像，增加搜索的精度。对于给定的颜色空间，其合并示意图如图 4-1 所示。

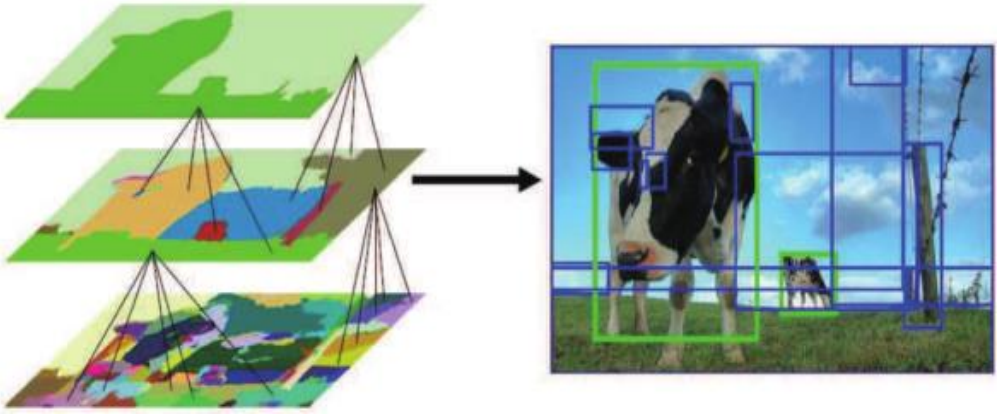


图 4-1 Selective Search 的分层结构[31,32]

给定一个超像素对(r_i,r_j)，其相似性可以通过公式(4-1)度量：

$$d(r_i,r_j)=a_1\cdot d_c(r_i,r_j)+a_2\cdot d_t(r_i,r_j)+a_3\cdot d_s(r_i,r_j)+a_4\cdot d_f(r_i,r_j) \tag{4-1}$$

其中d_c，d_t，d_s和d_f为四个基准的测量方式，分别表示倾向于优先合并颜色相似、纹理相似、小面积和有环绕现象的超像素对，$a_i\in\{0,1\}$表示是否选择该基准度量作为$d(r_i,r_j)$的一部分。以颜色相似性为例，计算任两个相邻的超像素的距离构成初始距离矩阵，然后合并距离最小的超像素对，将这个合并区域的最小方框包络输出为一个候选区域，并更新该区域与其它超像素的距离。迭代直至将所有超像素合并在一起。通过不同的度量方式组合，可以得到 16 种合并结果。

为了进一步增强搜索的空间范围，Selective Search 采用了不同的超像素初始化方式：1) 采用不同颜色空间：HSV 空间、Lab 空间、归一化 RGB 的 RG 通道加上亮度空间、HSV 中的 H 空间和亮度空间；2) 设置不同的初始化超像素大小，分别设置为 50，100，200 和 500。

组合合并会产生大量冗余窗口，后处理时会通过删除同样窗口的方法初步去冗余。Selective Search 的优点是其窗口定位准确性很高，因为它是通过非常小的超像素合并方式获得的。当需要从这些冗余的窗口中取一个很小的子集以提高后续分类器计算效率时，由于无法给出每个窗口含有目标的可能性，Selective Search 直接将所有窗口伪随机排序取出排序靠前的小部分窗口。显然这种取法不可靠。

4.1.2 EdgeBoxes

超像素合并过程需要大量的计算时间来初始化和更新权重矩阵，因此还有一类方法就直接通过多长宽比多尺度扫窗获得大量的窗口，然后计算每个窗口包含目标的置信度。当需要取出一定数量的窗口时，只需要按置信度排序取出得分靠前的部分窗口，取出的子集还能保持较高的召回率。这种方式虽然计算速度快，而且每个窗口都有置信度，但其窗口定位精确度却不如超像素合并的方式高。EdgeBoxes 是度量窗口包含窗口置信度的最典型的的方法，其原理图如图 4-2 所示。从上到下分别为：给定的图像，计算每个像素在边缘点上的概率与方向，通过概率和方向合并边缘点为完整的边缘，当窗口包含感兴趣目标时窗，口内完整

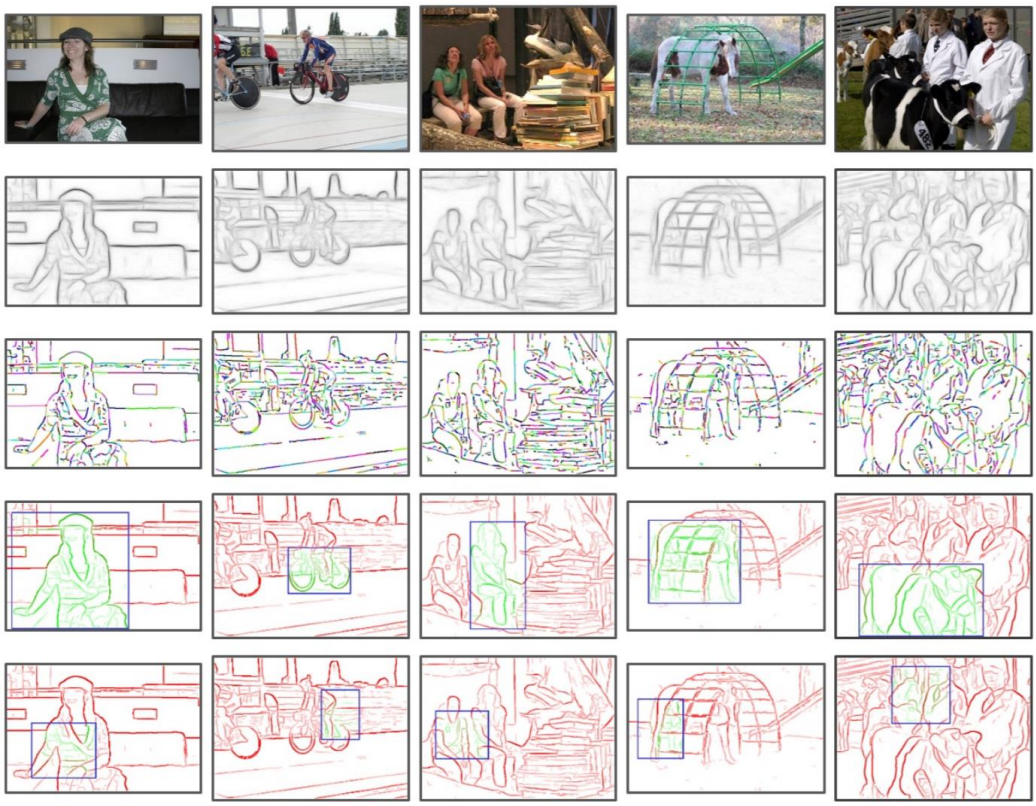


图 4-2 EdgeBoxes 原理示意图[34]

边缘的条数比较多，当窗口为一个背景区域时其含有的完整边缘条数比较少。

4.2 基于贝叶斯得分重排序的候选区域提取

正如前文所述，基于超像素合并的方式无序但定位精确，基于置信度的方式定位不精确但是有序，其特点互补。本文基于互补性提出了基于贝叶斯得分的重排序的候选区域提取方式，其示意图如图 4-3 所示。对于一幅输入图像，通过多分支侧输出网络同时获得边缘响应图和对称性响应图，使用改进的超像素合并方式获得冗余度很高的候选区域，并利用边缘和对称性信息分别计算每个候选区域在不同测度下的置信度，采用贝叶斯准则将不同测度的置信度合并为一个概率形式，基于概率对冗余候选区域进行重新排序。这种重排序策略只需要较少的时间代价就同时拥有了超像素合并方式与置信度方式的优点。

本节首先详细介绍改进的超像素合并策略相似性自适应搜索(Similarity Adaptive Search)，然后介绍基于图像底层特征的置信度计算，最后给出贝叶斯得分的重排序方法。

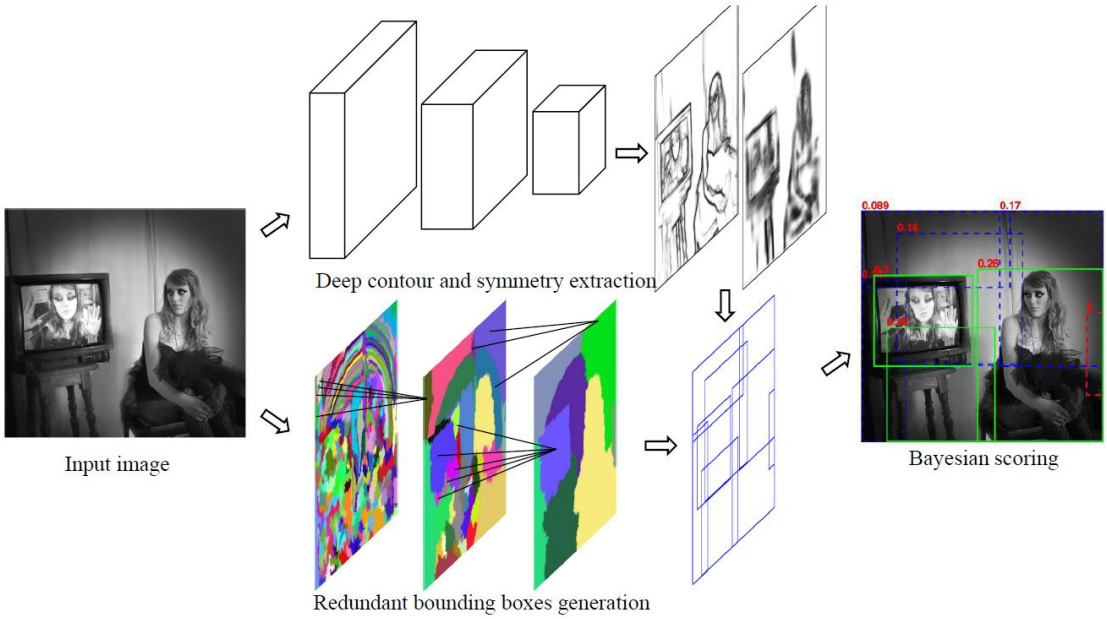


图 4-3 基于贝叶斯得分重排序的候选区域提取框架图

4.2.1 相似性自适应搜索

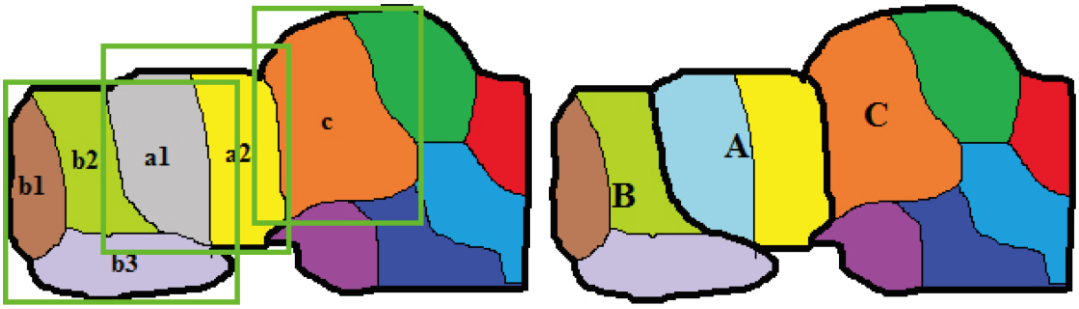
超像素合并的示意图如图 4-4 所示，超像素组$\{a_1,a_2\}$，$\{b_1,b_2,b_3\}$以及$\{c_1,c_2\}$通过合并策略合并之后输出其最小包络，如图 4-4(a)中绿色方框所示，其得到的新的超像素区块如图 4-4(b)所示。

对于新的超像素块，当考虑 Selective Search 中的颜色和纹理这两个测度时，其相似性为：

$$D_{mean}(R_m,R_n)=d_c(R_m,R_n)+d_t(R_m,R_n) \tag{4-2}$$

其中R_m和R_n表示超像素块，如图 4-4(b)中的区域A，B和C。从公式(4-2)可以看出，对于多个区域构成的超像素块是否合并是由整个区块的颜色和纹理的相似性决定的。

在很多时候，整个超像素区块的颜色是高复杂(high-complexity)的，也就是说平均颜色或者纹理并不能代表整个区块中的每个子超像素。如图 4-4(b)中，A区域应该和B合并还是和C合并，不仅取决于$d_c(A,B)$，而且还应该取决于相邻子超像素的距离$d_c(a1,b2)$和$d_c(a2,c)$。考虑到相连子超像素的低复杂度和高复杂度区域相似性被定义为[114]：



(a) (b)

图 4-4 超像素合并示意图

$$D_L(R_m,R_n)=\min\{d_c(r_i,r_j)+d_t(r_i,r_j)\,|\,r_i\in R_m,r_j\in R_n\} \tag{4-3}$$

$$D_H(R_m,R_n)=\max\{d_c(r_i,r_j)+d_t(r_i,r_j)\,|\,r_i\in R_m,r_j\in R_n\} \tag{4-4}$$

在相似性自适应搜索(Similarity Adaptive Search)中，两个超像素的距离被定义为：

$$\begin{aligned}D(R_m,R_n)=&b_1D_{mean}(R_m,R_n)\\&+b_2(\rho_{m,n}D_L(R_m,R_n)+(1-\rho_{m,n})D_H(R_m,R_n))\\&+b_3D_s(R_m,R_n)+b_4D_f(R_m,R_n)\end{aligned} \tag{4-5}$$

其中 $b_i\in\{0,1\}$ 表示是否选择该度量，$\rho_{m,n}$ 表示超像素组的动态指标，D_s 和 D_f 的定义和公式(4-1)中一样。当 R_m 和 R_n 中只含有一个超像素时，公式(4-5)和公式(4-1)一致。

本文在相似性自适应搜索之后，还采用了 MTSE(Multi-Thresholding Straddling Expansion)[115]进一步修正候选区域的定位精度。在[115]中，定义超像素紧致性为：

$$T=\sum_{s\in\mathcal{S}_b}\frac{|s|\,\delta(|s|-|s\cap b|)}{|b|}, \tag{4-6}$$

其中 $\mathcal{S}_b$ 为窗口 b 中的超像素的集合，$|\cdot|$ 表示定义区域的所有像素，$\delta(\cdot)\delta(|s|-|s\cap b|)$ =1 当且仅当超像素完全位于窗口 b 中。如果所有超像素都完全位于 b 内时，紧致性函数值为 1。但总有部分超像素有部分区域在 b 外，当扩大窗口使这些像素完全进入新的窗口内时，可以计算得到新的紧致性值。通过优化 T 使其最大化就能达到改善目标候选区域的目的。

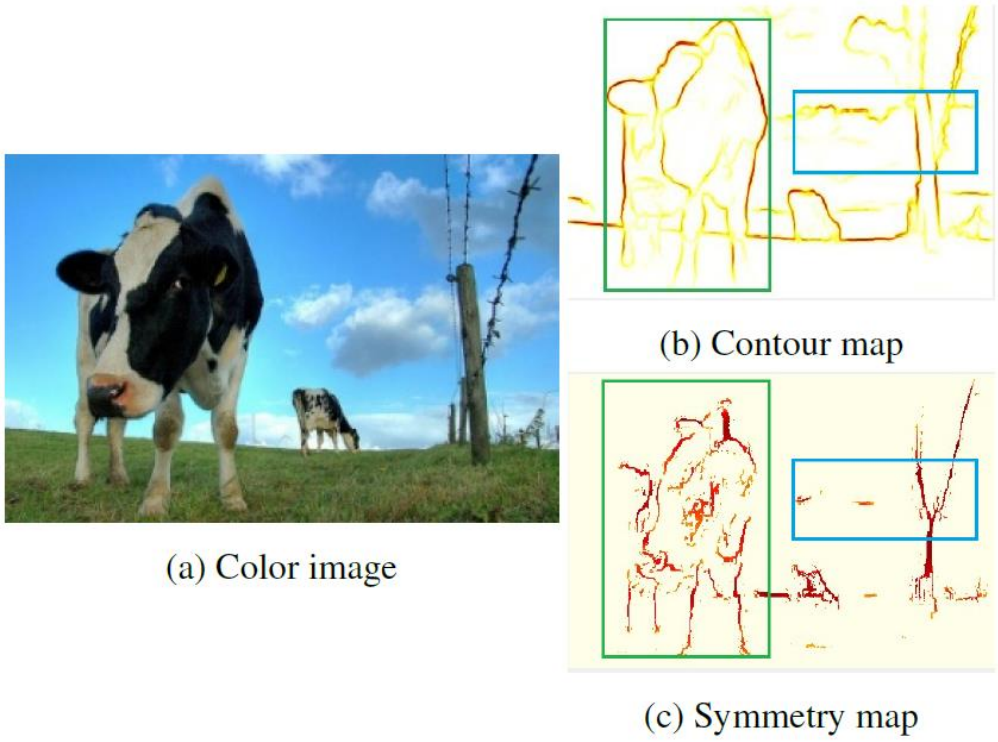


图 4-5 图像的底层边缘和对称性信息提取结果

4.2.2 候选区域置信度计算

根据第三章的相关描述，基于多分支侧输出残差网络结构可以同时快速准确地获得图像的底层边缘和对称性信息，如图 4-5 所示。可以从图 4-5(b)和图 4-5(c)中看到，含有目标的绿色窗口区域比没有目标的蓝色窗口区域所包含的边缘和对称轴的条数都会多一些。

边缘得分：采用[34]中的假设，如果一个窗口中包含的完整边缘条数越多，则其包含感兴趣目标的置信度就越大。给定一幅边缘响应图像，如图 4-5(b)所示，假设完整边缘的集合为$S=\{s_i\}$，那么候选框中的边缘组集合满足$S_b \subset S$，另外设完整边缘s_i上的点为$T=\{t_{ij}\}_{j=1}^{|s_i|}$，其中$|s_i|$表示像素个数。候选区域$b$的边缘得分可通过下式计算：

$$w_e=\frac{\sum_i w_b(s_i)m_i}{2(b_w+b_h)^{\kappa}}, \tag{4-7}$$

其中m_i是在完整边缘s_i上所有边缘点的模值之和，b_w和b_h是候选框的宽度和高度。采用周长进行归一化，归一化参数$\kappa>1$用来抵消更大的窗口包含更多完整边缘

引起的误差。边缘s_i的权重$w_b(s_i)$定义为：

$$w_b(s_i)=\begin{cases}1 & \text{if } s_i \text{ is in } b\\ 1-\max\limits_{P}\prod\limits_{j=1}^{|P|-1}a(t_j,t_{j+1}) & \text{if } s_i \text{ overlap } b\\ 0 & \text{if } s_i \text{ is out of } b\end{cases}, \tag{4-8}$$

其中$a(t_j,t_{j+1})$是边缘点t_j和t_{j+1}方向相似性度量，P为s_i上的有序序列，其长度为$|P|$。公式(4-8)表示当完整边缘完全位于窗口b中时，其权重为 1；当其完全位于窗口b外时，权重为 0；当其跨在窗口b上时，其权重由方向没有大角度变化的序列决定。

对于给定的边缘响应图，通过公式(4-7)计算窗口b的边缘得分的计算代价非常小，几乎可忽略不计。边缘得分越大，说明窗口b更有可能包含感兴趣目标物体。

对称性得分：对称性信息也是很重要的底层像素信息，类似于边缘得分的假设，当一个候选区域包含更多的完整对称轴时，其有更大的概率包含一个感兴趣物体。当给定一个对称性响应图，如图 4-5(c)所示时，可以通过类似于(4-7)的方式计算获得对称性得分。

4.2.3 候选区域重排序

通过相似性自适应搜索方式获得的冗余候选区域记做$B=\{b_1,b_2,\cdots,b_N\}$，每个窗口的边缘得分和对称性得分记做$H=\{(w_e^j,w_s^j)\}_{j=1}^N$。在贝叶斯框架下，窗口的最终置信度得分$y=f(H,B)$可以通过一个概率模型获得：

$$p(y|\mathcal{D}_N)\propto p(\mathcal{D}_N|y), \tag{4-9}$$

其中$\mathcal{D}_N=\{(H_j,b_j)\}_{j=1}^N$。边缘和对称性是两个独立的底层像素特征，因此可以假设其得分不相关，公式(4-9)可以写作：

$$p(\mathcal{D}_N|y)=\prod_{i=1}^{2}P(\mathcal{D}_N^i|y), \tag{4-10}$$

其中$\mathcal{D}_N^1=\{(w_e^j,b_j)\}_{j=1}^N$，$\mathcal{D}_N^2=\{(w_s^j,b_j)\}_{j=1}^N$。由于边缘得分和对称性得分是一个大于 0 的浮点数，使用$sigmo$函数将其变为概率形式，即$P((w_e,b)|y)=sigmoid(w_e)$和

$P((w_s,b)|y)=sigmoid(w_s)$。当给定一个窗口 b 以及其得分 $h=(w_e,w_s)$ 时，窗口中含有感兴趣物体的概率为：

$$p(y|(b,h)) \propto P((w_e,b)|y)\cdot P((w_s,b)|y). \tag{4-11}$$

4.3 实验结果及分析

实验中，首先评价了相似性自适应搜索、基于贝叶斯得分重排序在候选区域提取时的作用，然后将其与现有的方法做了详细地对比和分析，最后给出了行人数据集的评测。

4.3.1 实验配置

数据集：和候选区域提取的综述文章[30]一致。 在方法验证和对比中，我们使用 PASCAL VOC 2007 数据集[116]进行评测。PASCAL 数据集中有 20 类目标，包含了 4501 幅训练图像、2510 幅验证图像和 4952 幅测试图像。由于只关注无监督的候选区域提取方法，所以并不需要使用训练集。在验证自适应搜索和贝叶斯得分重排序的作用时，采用验证图像，在和其它方法比较时采用测试图像。除此之外，也评估了 MS COCO 数据集[6]的性能。COCO 数据集包含 80 类目标，40504 幅测试图像，更具有一般性。

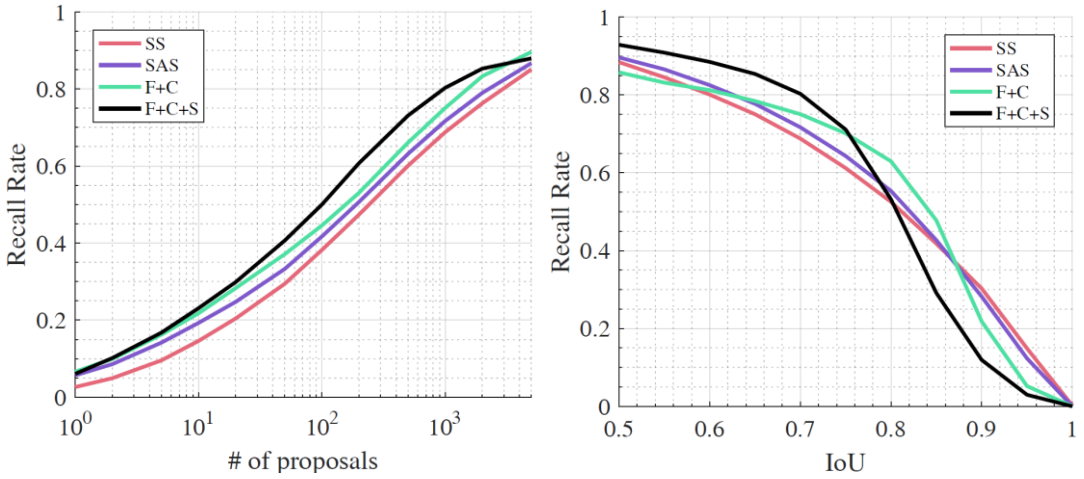
评测标准：候选区域提取的主要评测标准有召回率、窗口数以及交并比(Itersection over Union, IoU)[30,34]。

召回率：召回率是三个指标中最重要的标准。如果一些目标无法在候选区域中被选择出来，即使检测中下一阶段的分类器分类准确率再高，也无法检测出该目标。召回率越高则意味着检测器可能获得更高的检测率。

窗口数：窗口数量少，则分类的计算代价就比较小。而且窗口数量越少，可以使弱监督或者无监督的分类器的学习收敛得更快[31]。

交并比：交并比越大则说明候选区域的定位越精确，也可以使得检测器的学习越准确。

这三个指标是相互关联的，总体目标是在保证召回率的情况下，尽可能地降低窗口数和增加交并比。但三个指标一起评估很难实现，因此常用的评估设置为：给定交并比，不考虑窗口数的情况下能获得的最大召回率；给定交并比的情况下，



(a) 召回率和窗口数　　(b) 召回率和交并比

图 4-6 相似性自适应搜索和贝叶斯排序的有效性

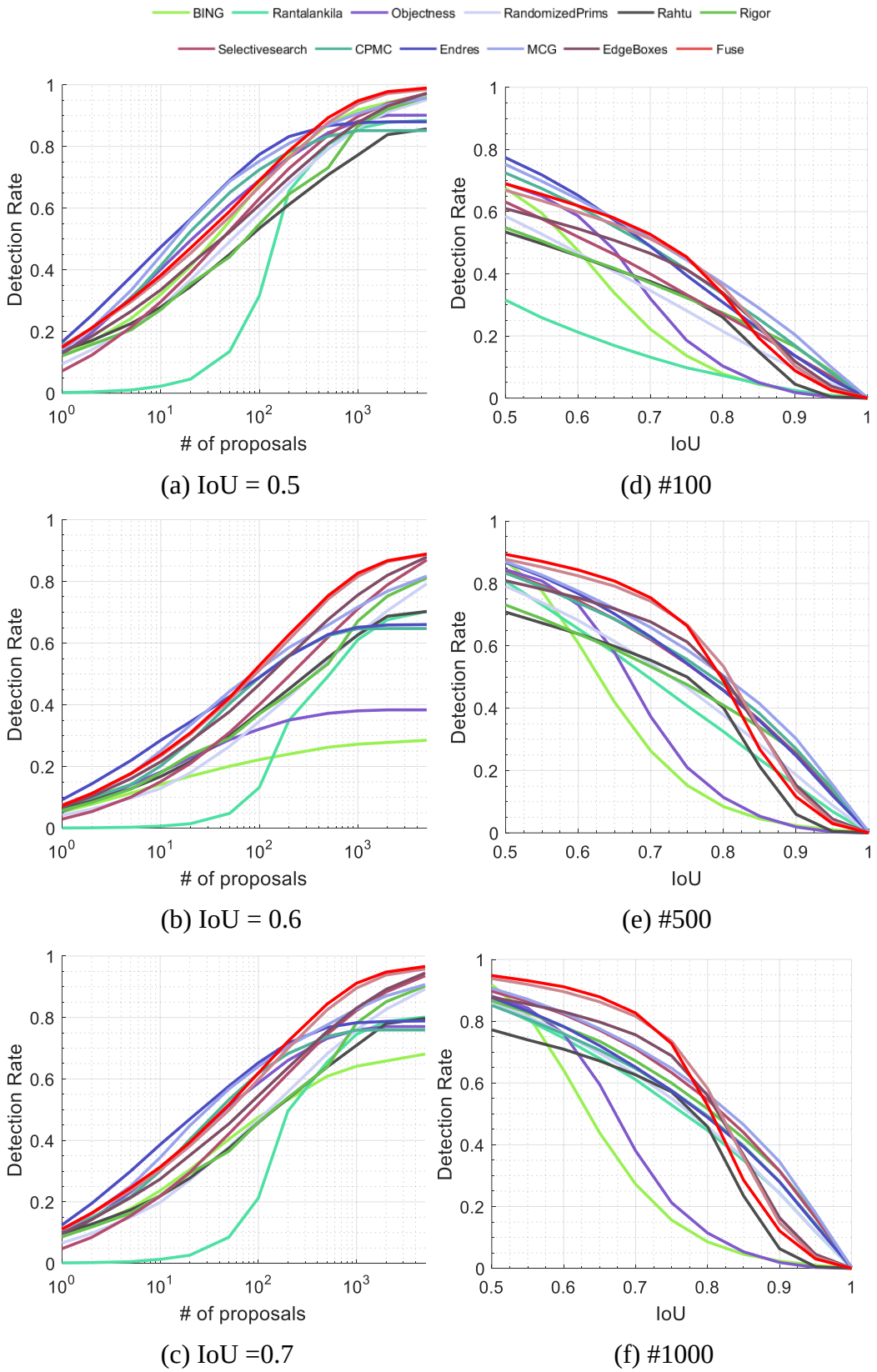
召回率和窗口数的关系；给定窗口数的情况下，召回率与交并比的关系。

4.3.2 与基准方法的比较

由于相似性自适应搜索是在 Selective Search[31,32]的基础上设计的，因此首先比较这两种方法的性能，然后比较只使用边缘得分做重排序(F+C)和同时使用边缘和对称信息(F+C+S)基于贝叶斯得分重排序的结果，如图 4-6 所示。

当给定 IoU=0.7 时，召回率与窗口数曲线如图 4-6(a)所示。通过黑色曲线和红色曲线的比较，可以看到，当使用 100 个窗口或者 1000 个窗口时，基于贝叶斯得分重排序的性能比 Selective Search 提高了 10%以上。总召回率也由 0.85 提高到了 0.89。基准方法 Selective Search 需要基于 1777 个检测窗口才能达到 75%的召回率，而基于贝叶斯得分重排序的方法只需要 601 个窗口。

选取 1000 个候选窗口时，召回率与交并比曲线如图 4-6(b)所示。从图中可以看出，当 IoU 在 0.5 到 0.8 之间时，基于贝叶斯得分重排序的方法要好于基准方法。当 IoU 大于 0.8 时，其召回率比基准方法低。造成这种情况的原因是在给每个窗口一个置信度之后，我们使用了非极大值抑制进一步降低了冗余度。由于交并比在 0.5 到 0.8 这个区间时，检测的定位已经非常准确了，所以基于贝叶斯得分重排序的候选区域提取方法综合性能要好于基准方法 Selective Search。



(a) IoU = 0.5　(d) #100

(b) IoU = 0.6　(e) #500

(c) IoU =0.7　(f) #1000

图 4-7 基于贝叶斯得分重排序的候选区域提取方法与现有方法的比较

Approach	AUC	N@25%	N@50%	N@75%	Recall
BING[33]	0.20	302	-	-	0.28
Rantalankila[124]	0.25	146	520	-	0.70
Objectness[117]	0.27	28	-	-	0.38
RandomizedPrims[122]	0.35	42	358	3204	0.79
Rahtu[123]	0.36	29	310	-	0.70
Rigor[121]	0.38	25	367	1961	0.81
Selective Search[31,32]	0.39	29	210	1416	0.87
MTSE[115]	0.41	18	175	1112	**0.89**
CPMC[119]	0.41	15	112	-	0.65
CA[114]	0.42	27	167	1418	0.88
Endres[120]	0.44	**07**	112	-	0.66
MCG[118]	0.46	10	86	1562	0.82
EdgeBoxes[34]	0.47	12	96	658	0.88
Our approach (C)	0.48	12	91	535	0.88
Our approach (C+S)	**0.49**	10	**71**	**476**	**0.89**

表 4-1 取得 25%，50%和 75%召回率时需要的最少候选区域个数

4.3.3与其它方法比较

在 PASCAL VOC 2007 的测试集上，将基于贝叶斯得分重排序的候选区域提取方法和其它数十种方法进行了比较，分别是 Objectness[117]、MCG[118]、CPMC[119]、BING[33]、Endres[120]、Rigor[121]、RandomizedPrims[122]、Rahtu[123]、Rantalankila[124]、Selective Search[31,32]、Complexity-Adaptive(CA)[114]、MTSE[115] 和 EdgeBoxes[34]。除了本文方法之外，其它方法的数据均从[30]获得，评测代码使用 EdgeBoxes 的评测代码[34]。

分别给定交并比(IoU)为 0.5、0.6 和 0.7 时，召回率与窗口数的曲线图如图 4-7(a)到图 4-7(c)所示。如图中红色曲线所示，无论采用何种交并比，基于贝叶斯得分重排序的候选区域提取方法的最大召回率均高于其它方法。分别设定候选区域窗口数为 100、500 和 1000 时，召回率与交并比的曲线图如图 4-7(d)到图 4-7(f)所示。图 4-7(d)中给定了 100 个窗口，当 IoU 在 0.5 到 0.7 之间时，Endres, CPMC 和 MCG 要略好于本文提出的方法。当给定 500 个或者 1000 个窗口时，本文的方法达到了最好的性能。

在表 4-1 中，比较了各种方法在给定 IoU=0.7 不限制窗口数量时的最大召回率。以及 1000 个窗口时的 AUC(Area under Curve)值。贝叶斯得分重排序的候选

区域提取方法和 MTSE 同时达到了最佳的召回率 0.89，比基准方法 Selective Search 高了 0.02。当只使用边缘得分时，AUC 从 Selective Search 的 0.39 增加到 0.48，使用贝叶斯得分重排序之后进一步增加到 0.49。表 4-1 还比较了在分别获得 25%，50%以及 75%的召回率时，各个方法所需要的最少窗口数量。75%的 recall 对于目标检测来说是一个比较好的权衡值。在这个召回率下本文提出的方法只需要 476 个窗口，比其它方法都要少。尤其是加入重排序之后，比伪随机排

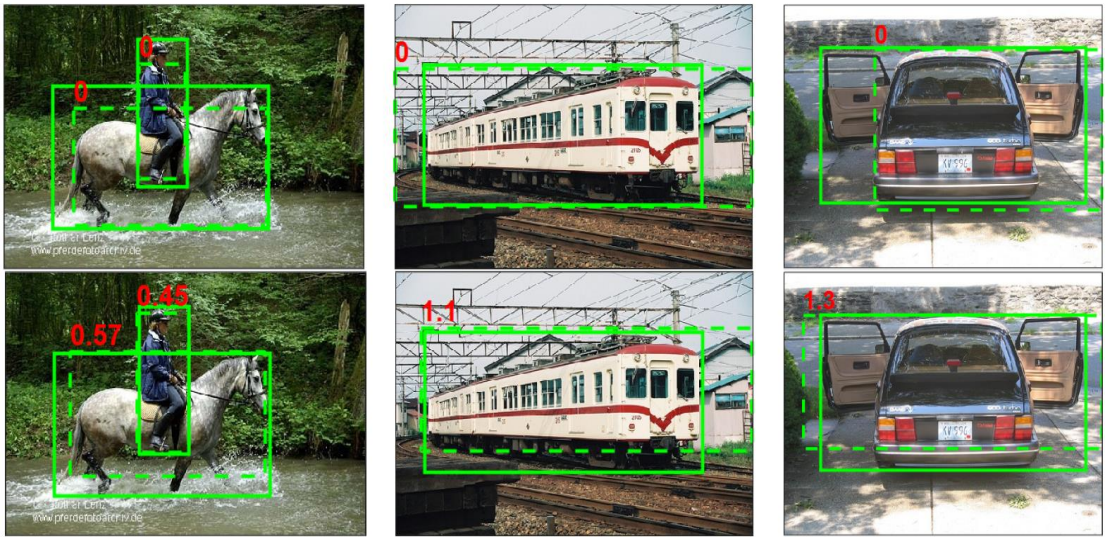


图 4-8 候选区域目标定位准确性比较

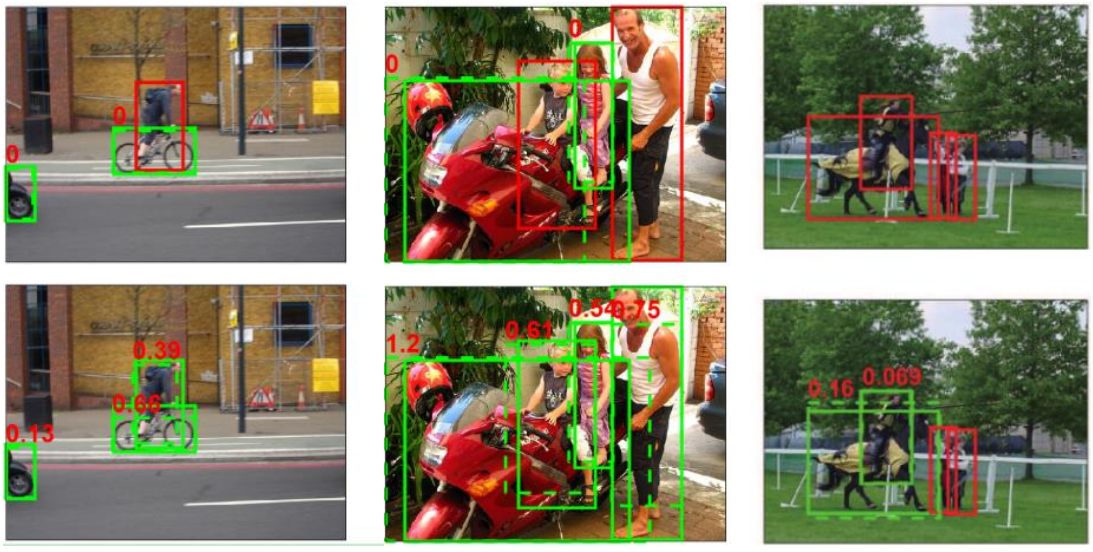


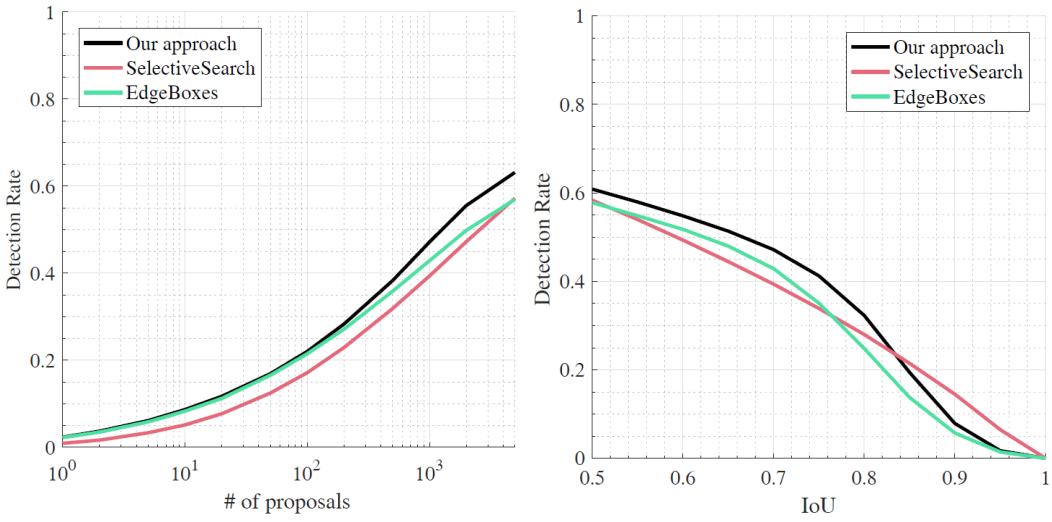
图 4-9 候选区域目标召回率比较

序的基准方法 Selective Search 要少将近 1000 个窗口。

使用 1000 个窗口的情况，IoU 限制大于 0.7 时，基于贝叶斯得分重排序的候选区域方法与 Selective Search 方法的候选区域的结果分别如图 4-8 和图 4-9 的第一行和第二行所示。图 4-8 展示了定位准确性的比较，本文的方法通过排序之后将准确性更高的区域提前。图 4-9 展示了召回率的比较，本文的方法通过排序将更多的含有目标的候选区域提前。

4.3.4 COCO 数据集实验结果及分析

COCO 数据集因其含有更多的目标以及更多的测试图像而更具有挑战性。实验只比较了本文的方法和两类候选区域提取方法中最经典的 Selective Search[31,32] 和 EdgeBoxes[34]，结果如图 4-10 所示。虽然 COCO 相比于 PASCAL VOC 更具有挑战性导致了所有方法的最大召回率下降到 60%左右，本章的方法仍然从 Selective Search 的 57%提升 6 个百分点到 63%。



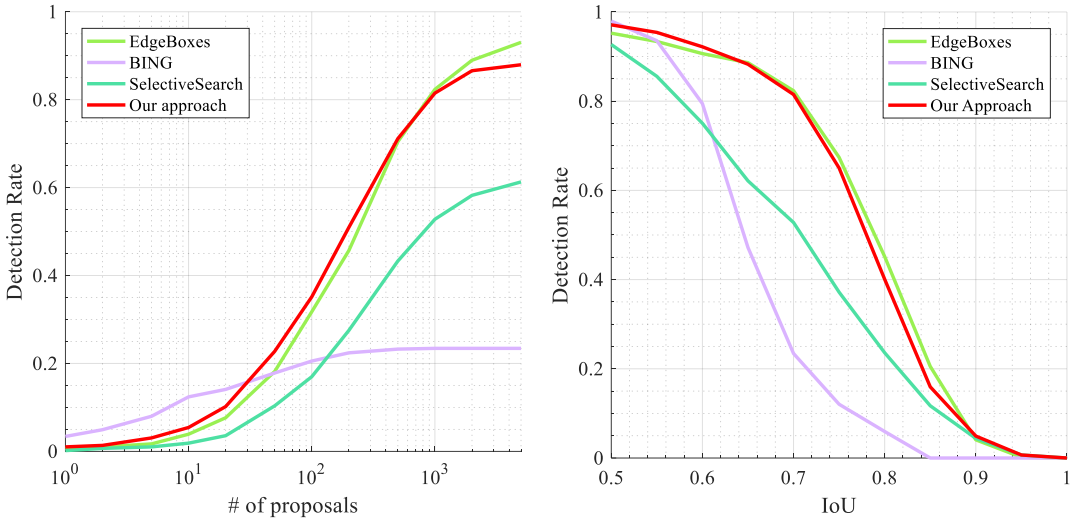
(a) 召回率和窗口数　　(b) 召回率和交并比

图 4-10 COCO 数据集上候选区域性能比较

4.3.5 行人数据集上的实验结果及分析

在行人数据集 INRIA 上，基于贝叶斯得分重排序的候选区域提取方法与其它方法的结果如图 4-11 所示。当 IoU=0.7 时，当使用的窗口数在数百的量级时，本章方法的召回率最高，体现了降低窗口数量的有效性。当采用 1000 个窗口时，

本章的方法在 IoU 在 0.50 到 0.65 时性能略好于 EdgeBoxes，而其它取值时略差于 EdgeBoxes。对于行人这种单目标而言，现有的非监督候选区域提取方式的的召回率性能还远达不到采用弱分类器提取的候选区域的的性能，这将在第五章第三节中详细讨论。



(a) IoU=0.7　　(b) #1000

图 4-11 INRIA 数据集上候选区域提取的比较

4.4 本章小结

本章结合超像素合并和置信度两类候选区域提取方式的优点，提出了基于贝叶斯得分重排序的候选区域提取方法。冗余候选区域由相似性自适应搜索获得，改进了超像素集合之间的相似性度量。置信度则是在贝叶斯框架下将对称和边缘信息进行统一，衡量冗余候选区域中包含感兴趣目标的概率。实验表明，选择高置信度的候选区域子集，确实可以在保持召回率和定位准确性的同时，大幅度降低候选区域的个数。候选区域个数地降低，在全监督检测中可以节省分类时间，在自学习分类器中可以降低样本搜索空间。

第 5 章 基于 PCA 卷积特征的全监督行人检测

自 R-CNN[13]的出现，目标检测算法全面从扫窗转换到首先进行候选区域提取，再进行精细分类的框架上来。即使在最新的目标检测方法中，进行端到端的训练，也仍然需要这两个模块的存在[25-27]。

基于 PCA 卷积特征的全监督行人检测方法中，在特征设计上借鉴了卷积神经网络，但是通过固定的卷积核进行卷积运算以增加特征的表达能力，仍属于手工设计特征的一种；在框架上借鉴了 R-CNN 的思想，先获得一些候选区域之后再做精细分类，因此重点关注于精细分类步骤中的特征设计。本章中，首先介绍了基准方法通道特征[9]，然后详细介绍了基于 PCA 卷积特征的全监督行人检测方法，最后在实验中分析了行人检测中候选区域提取方式以及最终的检测结果。

5.1 基于聚合通道特征的行人检测简介

当给定一幅图像 I，通道特征(ACF, Aggregate Channel Feature)通过多个线性或者非线性的变换函数将图像映射到一个新的空间。由于这个映射函数是分别作用于各个像素的，因此特征空间中仍然是一幅一维图像，将其命名为通道特征。映射函数可以表示为：

$$f = \Omega(I) \tag{5-1}$$

Ω 的选取多种多样，如图 5-1 所示，可以是单个像素处理的颜色空间的变换，也可以是以很小的局部小块提取的梯度和边缘；可以是多尺度高斯差分(DoG, Difference of Gaussian)，也可以是局部区域的统计梯度直方图(Gradient Hist)。当

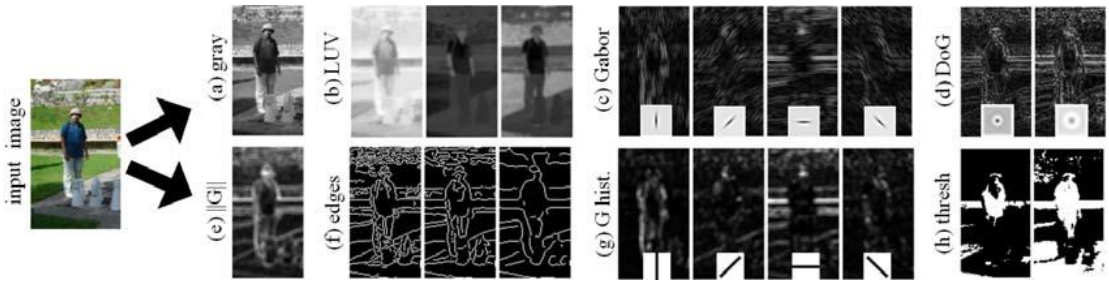


图 5-1 行人样本的通道特征

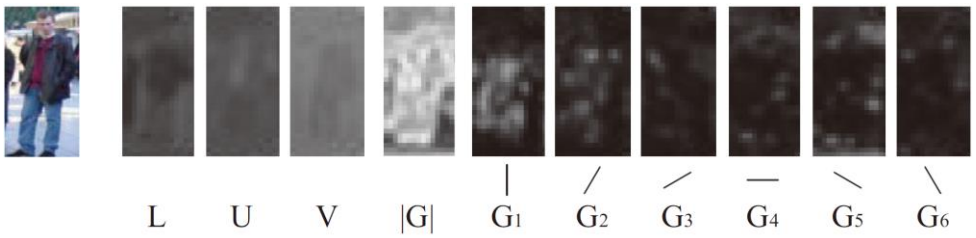


图 5-2 行人样本的较优通道特征

Ω 是一个一阶函数时，可以作用在单独的像素点上，例如取出某个颜色分量的值，如图 5-1(a)、5-1(b)和 5-1(h)所示，也可以作用以某个像素为中心的一小片区域上，例如水平方向或者竖直方向的梯度值，如图 5-1(e)和 5-1(f)所示。当 Ω 为一个高阶函数时，它为多个一阶函数的组合，会提取小片区域的特征值，如图 5-1(c)、5-1(d)和 5-1(g)所示。在 ACF 中，通过实验选取出了效果比较好的十个通道，分别是：LUV 空间中的三个颜色通道、归一化的梯度的模值（|G|通道）、梯度六个量化方向（G1-G6），如图 5-2 所示。

传统的多尺度扫窗检测框架中，需要构建稠密尺度的特征金字塔以覆盖所有尺度的目标，这往往成为检测器最耗时的部分。而在使用通道特征，构建图像特征金字塔时，由于其具有很好的差值拟合特性[45]，可以大大降低特征提取时间。在构建图像金字塔时，只需要计算每个阶的特征，阶内的特征全都通过该阶第一层和最后一层特征进行拟合。为了进一步增加运算速度，还在 ACF 中对每个通道进行了 2×2 下采样。

虽然 ACF 的计算十分简单而且非常快速，但是在其特征空间中线性可分度却很差，不适合于使用 SVM 作为分类器。AdaBoost 分类器是 ACF 的最佳选择，如图 5-3 所示。一方面，AdaBoost 分类器可以学习到非线性的分界面，另一方面级联的 AdaBoost 分类器还可以起到特征选择的作用，将 ACF 的高维特征降低到很低的维度(等于 AdaBoost 级联的弱分类器的个数)。

在 ACF 中，两层的决策树被用作弱分类器，分别级联 32、128、512、2048 棵决策树构成四个阶段的分类器。在每个阶段后会进行一次反例样本挖掘以增加分类的准确性。也正是因为挖掘到了更难的反例样本，所以随着阶段的增加，

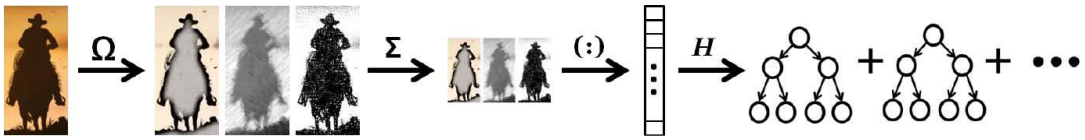


图 5-3 结构化决策树构成的 AdaBoost 分类器示意图

所需要级联的决策树的数量也在增加。最终的分类器是由第四阶段的 2048 棵决策树构成的强分类器。

5.2 基于 PCA 卷积特征的全监督行人检测

PCA 卷积特征通过 PCA 算法[125]获得卷积核，设计两层卷积网络将通道特征进行张量运算使其表达能力更强，同时正交投影运算使也其更符合 AdaBoost 分类器。

5.2.1 PCA 卷积核学习

在 PCANet[125]中，直接在原始图像是使用 PCA 算法获得卷积核，而在本章中，由于不同的通道特征之间相关性较弱，每个通道都学习了一组卷积核，如图 5-4 所示。

给定 N 个正例样本 $\mathrm{I}=\{I_i\}_{i=1}^{N}$，其中每个样本的大小为 $w\times h$。它们的 ACF 特征被记作 $\{f_{i,k}\}$，其中 $i=1,2,\cdots,N$ 和 $k=1,2,\cdots,K$，表示第 i 个样本的第 k 个通道特征。对于第 k 个通道，在每个像素周围提取一个 $m\times m$ 的局部快。将所有的局部块 $x_{j,k}$ 变成列向量之后，可以得到局部块矩阵：

$$\mathrm{X}_k=[X_{1,k},X_{1,k},\cdots,X_{M,k}]\in \mathrm{R}^{mm\times Mwh} \tag{5-2}$$

其行数为每个块中像素的个数 mm，其列数 M 为所有 N 个样本的像素总个数 Mwh。

PCA 算法的作用是将一系列可能相关的数据正交地投影到一些线性不相关的特征轴上。它的优化目标是在一组正交基上最小化重构误差。对于第 k 个通道特征，目标函数为：

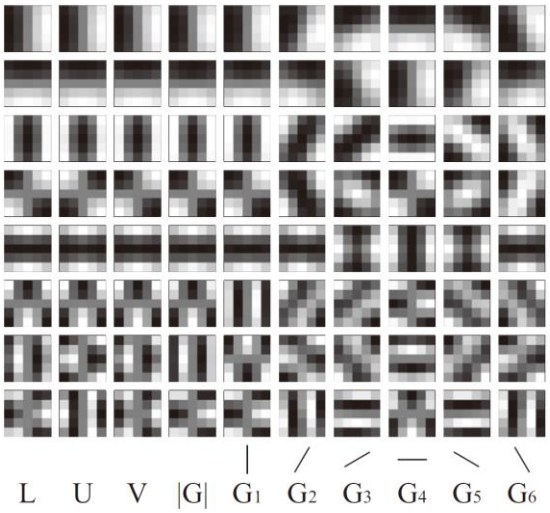


图 5-4 不同通道特征的 PCA 卷积核

$$\begin{aligned} &\min \left\| \mathrm{X}_k - VV^T \mathrm{X}_k \right\|^2 \\ &s.t.\ VV^T = I_L \end{aligned} \tag{5-3}$$

其中 I_L 为一个 $L \times L$ 的单位阵，以约束 V 中的基正交。通过求解公式(5-3)就可以获得 V，其每一列都为局部快矩阵的一个特征向量。将这些特征向量重新变为二维 $m \times m$ 的矩阵，就可以当做一组正交卷积核：

$$P_{l,k} = mat(V_i) \in \mathrm{R}^{m \times m}, l = 1, 2, \cdots, L \tag{5-4}$$

其中 $mat(\cdot)$ 表示从向量转换为矩阵的操作。

在实验室中，PCA 卷积核的大小被设置为 $m \times m = 5 \times 5$。由于 PCA 的特征向量是按照特征值的大小排序的，特征值越大特征向量含有的原信息就越多，因此每个通道上我们选取了前 L 个特征向量作为 PCA 卷积核。如图 5-4 中每列卷积核中第一个都包含了最多的原始信息，特别是六个方向梯度特征的第一个 PCA 卷积核呈现的模式与方向一致。

5.2.2 PCA 卷积特征提取

PCA 卷积特征由四部分构成，分别为聚合通道特征提取、卷积、空间池化以及通道间池化，如图 5-5 所示。其中 ACF 直接采用[9]中的十个通道特征，如图 5-5 中的 ACF extraction 所示。

卷积：每个通道特征中，使用 PCA 卷积核进行卷积，表示为：

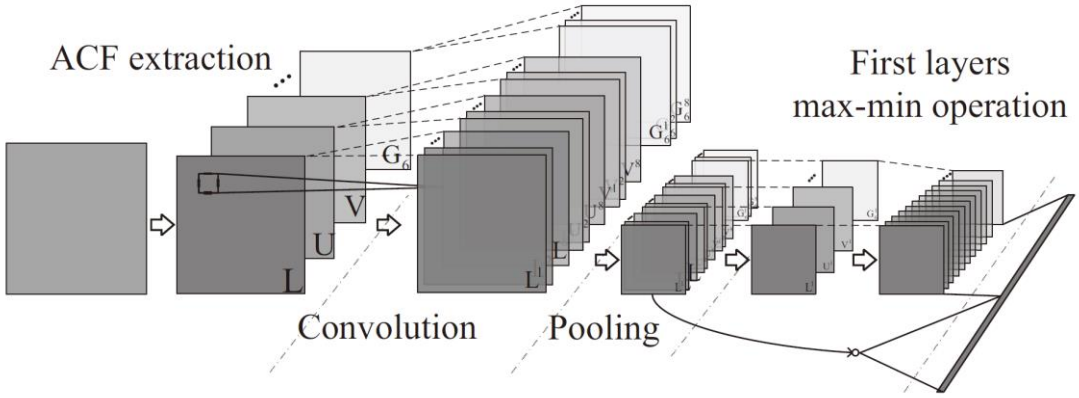


图 5-5 PCA 卷积特征提取

$$CONV_{l,k} = P_{l,k} \circ f_k \tag{5-5}$$

其中 $P_{l,k}$ 以及 f_k 为卷积核和第 k 个通道特征。经过卷积之后，一个样本的通道特征由原来的 K 增加到 $L \times K$ 个，进行正交分解的同时增加了特征的表示能力，如图 5-5 中的 Covolution 所示。

空间池化：类似于卷积神经网络，在 PCA 卷积通道特征中，也在特征层上进行池化，如图 5-5 中的 Pooling 所示。池化的好处在于可以降低特征的维度，也可以有效地保持局部信息的统计特性。因此在卷积层之后采用了典型的 2×2 块的最大及最小池化：

$$\begin{cases} FS\_MAX_{l,k}(i,j) = \max\{CONV_{l,k}(R(i,j))\} \\ FS\_MIN_{l,k}(i,j) = \min\{CONV_{l,k}(R(i,j))\} \end{cases}, \tag{5-6}$$

块的最大及最小池化：

$$\begin{cases} FS\_MAX_{l,k}(i,j) = \max\{CONV_{l,k}(R(i,j))\} \\ FS\_MIN_{l,k}(i,j) = \min\{CONV_{l,k}(R(i,j))\} \end{cases}, \tag{5-6}$$

其中 $CONV_{l,k}(R(i,j))$ 表示以第 k 个通道使用第 l 个 PCA 卷积核卷积之后的特征图上，以像素点 (i,j) 为中心的一个 2×2 的区域。由于在 ACF 中就使用了下采样，因此经过池化操作后的特征图 $FS=\{FS\_MAX, FS\_MIN\}$ 的维度为 $2\times L\times K\times W\times H/4$，其中 $W\times H$ 是输入图像的大小。

通道间池化：在 PCANet[125]中需要对每组卷积层进行合并以防止每个卷积层

单独处理出现信息丢失。类似地，也希望增加一些 ACF 十个通道之间的相互信息。将池化后的层任取两层进行通道间池化是一个可行的方法，这是因为层间最大池化类似于“或”操作，同时保留对应两层所有的信息，层间最小池化类似于“与”操作，只保留对应两层共有的信息，如图 5-5 中 First layer max-min pooling 所示。这种通道间池化需要进行 $2\times C_{10L}^{2}$ 次，为了防止维度灾难，只需在每一组中提取信息含量最大的卷积层进行通道间池化。通道间池化表示为：

$$\begin{aligned} FC\_MAX_{l,k}(i,j) &= \max\{FS_{1,k^{(1)}}, FS_{1,k^{(2)}}\} \\ FC\_MIN_{l,k}(i,j) &= \min\{FS_{1,k^{(1)}}, FS_{1,k^{(2)}}\} \end{aligned} \tag{5-7}$$

其中 $FS_{1,k^{(1)}}, FS_{1,k^{(2)}}$ 表示 $k^{(1)}$ 和 $k^{(2)}$ 两个通道经过卷积后的第一层特征响应图，也就是使用第一个 PCA 卷积核获得的卷积响应图。

最终的特征是将空间池化层以及通道间池化层合并成的长向量，表示为：

$$F = \{FS, FC\}. \tag{5-8}$$

5.2.3检测器实现

因为 PCA 卷积特征已经构成一个卷积网络的简单结构，其计算的复杂度远高于 ACF 本身，提取候选区域之后再对候选窗口进行检测的框架则更加合适，整个行人检测框架如图 5-6 所示。由于行人这类目标比较特殊，很多行人检测的方法都可以达到实时的处理速度。因此在提取候选区域时采用快速的弱分类器不仅可以达到目标候选区域高召回率、高定位精度、所需候选区域数量少和速度快的标准，而且弱分类器的置信度也可以对最终的判断进行指导。在提取候选区域时，只需降低 ACF 中的 AdaBoost 级联分类器的级数就可以有效地保持召回率，而且 ACF 中扫窗的使用可以保证定位准确性。

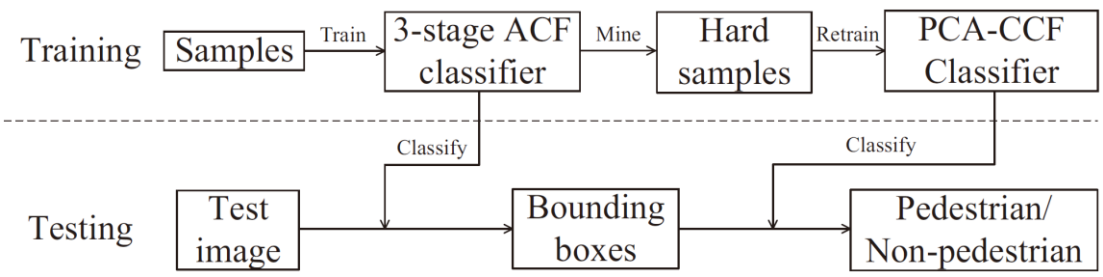


图 5-6 基于 PCA 卷积特征的行人检测流程图

如图 5-6 所示，在训练阶段，先训练一个三阶段 ACF 分类器，并用其进行难反例样本挖掘。然后用正例样本和挖掘到的难反例样本重新训练一个 PCA 卷积分类器。在测试阶段，对于一幅输入图像，使用三阶段 ACF 分类器获得一些候选区域，然后再用 PCA 卷积分类器进行更加精确的分类。

5.3 实验结果及分析

实验中，首先验证了使用 ACF 进行目标预定位要比上一章的目标预定位方法在行人这类目标上更有效，然后比较了 PCA 卷积特征各个组成部分的有效性，最后与其它行人检测方法进行了横向比较。

5.3.1 实验配置

数据集：采用公开数据集 INRIA[2]和 Caltech[126]进行评测。

INRIA 数据集在很长一段时间内都是做全监督行人检测评测的标准数据集。它是由单独拍摄的高清图像组成的，训练数据中包含 614 幅正例图像，1218 幅反例图像，测试数据中包含 288 幅图像。训练集中的正例图像有 1208 个行人样本，反例图像中不包含行人样本，主要用来挖掘反例样本。INRIA 数据集因为采用的的是单独拍摄的照片，因此其中的行人样本的分辨率一般都比较高，而且很少出现遮挡。

Caltech 数据是采集的 640*480 的交通场景视频，分辨率很低。标注了大约 250000 帧的图像，约含有 350000 个行人样本，其中 67000 幅训练图像，65000 幅测试图像。Caltech 中行人的尺度变化很剧烈，最大的行人样本高度有 97 个像素，而最小的只有 27 个像素，因此将其分为远、中、近三个子集，可以分别做评测也可以放在一起评测。除了这种分类方式之外，Caltech 中大于 50 像素的样本构成的 Reasonable 子集也经常被用来做评测。Caltech 的另一个特点是遮挡情况严重，除了行人之间的遮挡，还有行人被车辆、树木等遮挡一部分，这些都增加了检测的难度。

评测标准：评测使用[126]中的错检率(miss rate)与平均每幅图像的误警率(FPPI, False Positive)曲线，错检率和 FPPI 均在对数空间中，以便能将性能曲线更好地区分开。

对于一个检测窗口 BB_{dt} 是否正确的衡量标准是：

$$a_o = \frac{area(BB_{dt} \cap BB_{gt})}{area(BB_{dt} \cup BB_{gt})} > 0.5 \tag{5-9}$$

其中 BB_{gt} 表示正例样本的窗口，即 BB_{dt} 和 BB_{gt} 的交并比大于 0.5 时就认为检测是正确的。错检率 miss rate 则定义为 1-detection rate，即检测精度越高错检率越低。给定一个目标得分的阈值后，就可以计算该阈值下的错检率和 FPPI。当给定一系列阈值之后就可以获得错检率曲线。

由于曲线并不适于定量的衡量，因此还是用到了平均错误率。平均错误率为 FPPI 从10^{-2}到10^{2}中均匀分布的九个点的错误率的平均值。

5.3.2 候选区域评测

在全监督行人检测中，有监督信息可以使用，所以可以训练弱分类器进行候选区域提取。本小节比较了三阶段 ACF 弱分类器和上一章中的典型非监督方式 Selective Search[31,32]、BING[33]以及 EdgeBoxes[34]的候选区域提取性能，如图 5-7 所示。

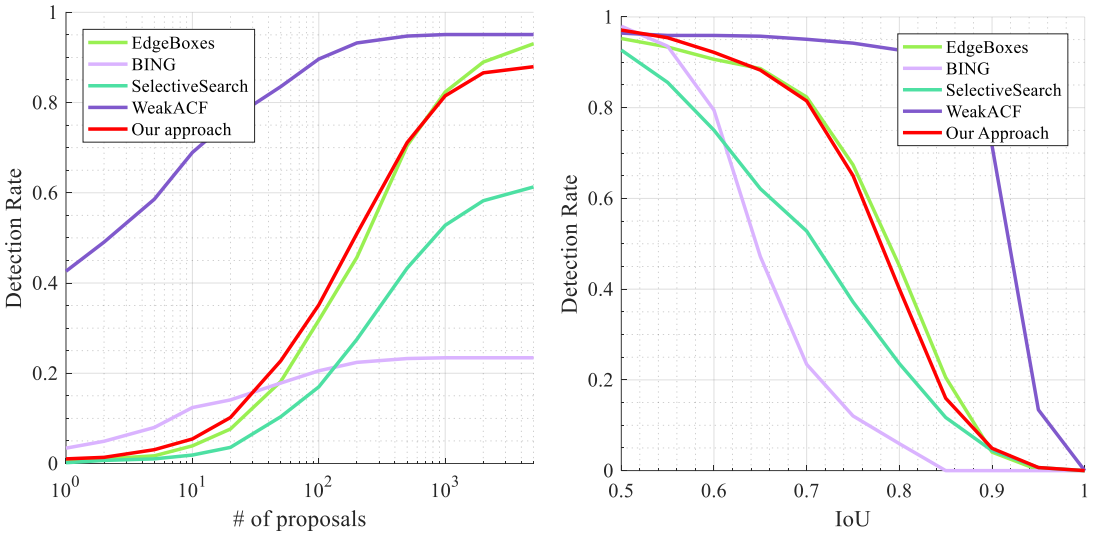
图 5-7(a)展示了在给定 0.7 的交并比的情况下，召回率与候选区域数量之间的关系。三阶段 ACF 弱分类器在使用 1 个候选区域时就能达到 0.40 的召回率，其它方法都不到 0.05；最大召回率可以达到 0.95，BING、Selective Search 和 EdgeBoxes 分别为 0.23、0.61 和 0.93；在使用 300 个窗口时，基本达到最大召回率，而 EdgeBoxes 需要 3000 个以上。

图 5-7(b)展示了在给定 1000 个窗口时，召回率与交并比之间的关系。上一章已经讨论过对于目标检测而言，交并比在 0.7 时已经定位相当准确。BING 在 IoU=0.7 时，召回率大幅度降低，Selective Search 也降低了三分之一以上。当 IoU=08 时，EdgeBoxes 的召回率降低到不足 IoU=0.5 时的一半。而三阶段 ACF 弱分类器即使在交并比达到 0.9 时仍然没有出现明显的下降，一直维持在很高的水平。

就运算速度而言，BING 最快，可以达到 100 帧每秒，三阶段 ACF 的运算可以达到 14.3 帧每秒，而 EdgeBoxes 处理一帧图像需要 0.2 秒左右，Selective Search

则需要 3 到 5 秒。三阶段 ACF 弱分类器基本已经达到实时。

综上所述，对于全监督的情况，采用三阶段 ACF 弱分类器提取候选区域除了运算速度略输于 BING，其它各个方面的指标均远远高于其它方法，因此三阶段 ACF 弱分类器成为全监督行人检测候选区域提取的最好选择。



(a) IoU=0.7　　(b) #1000

图 5-7 全监督行人候选区域提取性能比较

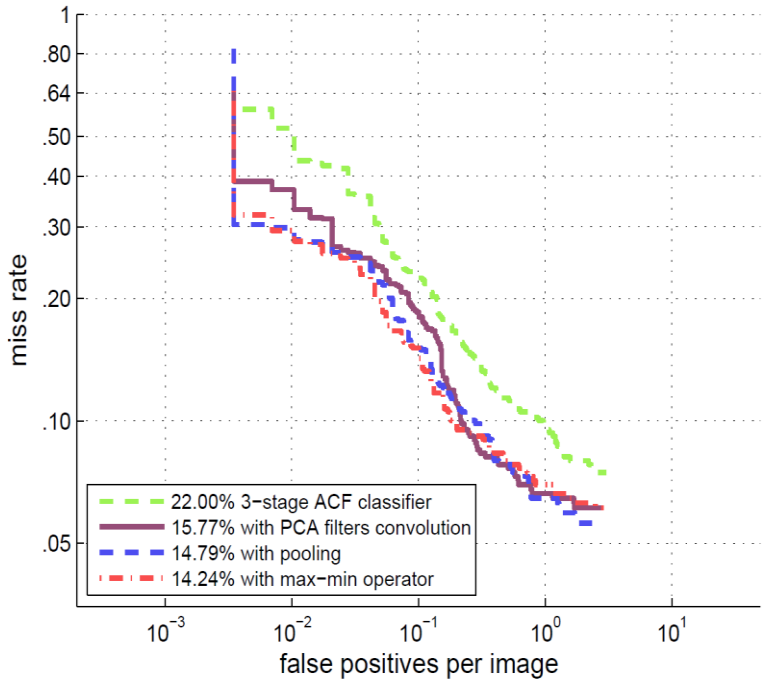


图 5-8 PCA 卷积特征有效性验证

5.3.3 PCA 卷积特征有效性验证

验证 PCA 卷积特征有效性的实验是在 INRIA 数据集上进行的，实验结果如图 5-8 所示。三阶段 ACF 弱监督分类器分别使用 32,128 和 512 棵决策树。PCA 卷积核数取 8，大小取 5×5。

弱分类器本身也可以作为行人检测器，其错检率为 22.00%。当只使用单层 PCA 卷积特征时，错检率为 15.77%，提升了 6.23%。当加入空间池化和通道间池化时，错检率分别为 14.79%和 14.24%，分别提升了 0.98%和 0.55%。在任何任务中，当性能趋于饱和时，很小的提高都是有意义的。因此图 5-8 证实了 PCA 卷积特征在各个阶段的有效性。

5.3.4 全监督行人检测实验结果及分析

本小节在 INRIA 和 Caltech 数据集上比较了基于 PCA 卷积特征的行人检测方法和其它方法的性能。这些方法分别是三种传统的方法 VJ[1]、HOG[2]和 LatSvm-V2[23,52]，基准方法 ACF[9]，和其它十种性能优秀的方法包括：Shapelet[53]、Chn-Ftrs[44]、ConvNet[48]、Sketch Tokens[49]、VaryFast[127]、FPDW[45]、Roerei[128]、

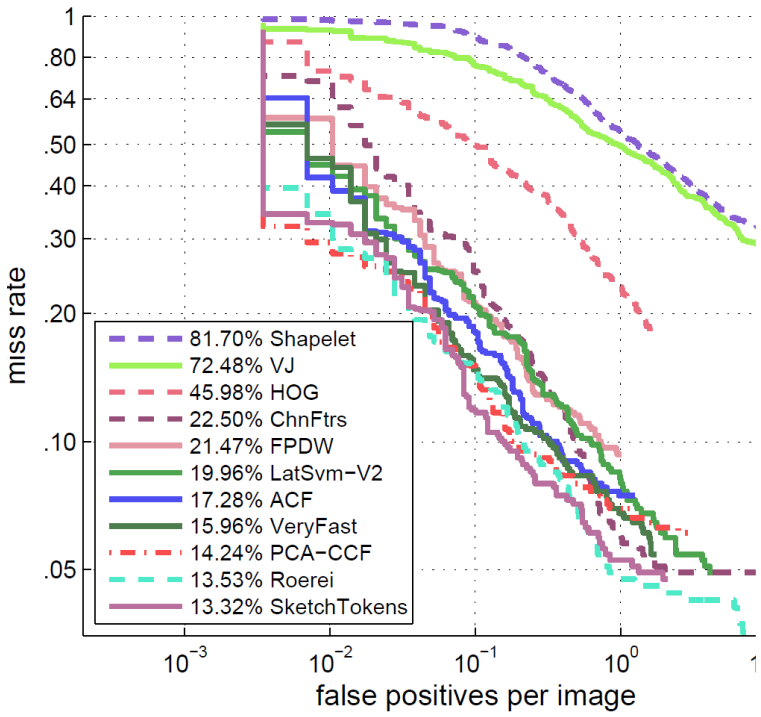
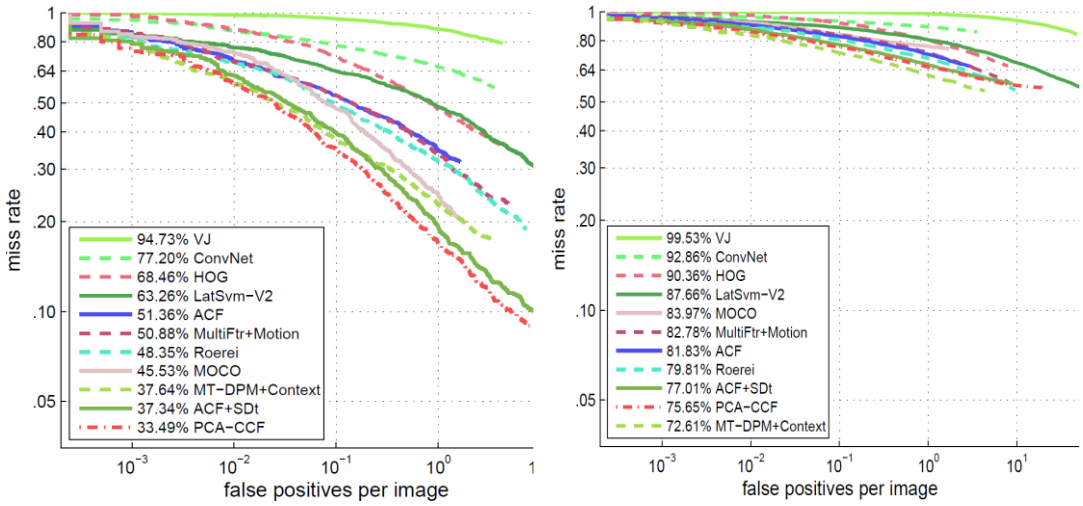


图 5-9 INRIA 数据集上的行人检测性能曲线



(a) Reasonable (b) Overall

图 5-10 Caltech 数据集上的行人检测性能曲线

MultiFtr-Motion[62]、MOCO[65]、MT-DPM+Context[59]和 ACF-SDt[63]，所有对比方法的检测曲线都通过[129]获得。图 5-9 展示了在 INRIA 数据集上的行人检测性能曲线。很多方法性能已经比经典的 VJ、HOG 高出 20%以上，处于一个新阶段里。在这个阶段中，PCA 卷积特征分类器性能为 14.24%，相比于基准方法 ACF 提高了 3.04%，与最好的 Roerei 和 SketchTokens 的性能已经非常接近，后两者的性能分别为 13.53%和 13.32%。

图 5-10 展示了 Caltech 数据集上的行人检测性能曲线。图 5-10(a)为 Reasonable 子集的检测结果。在高度大于 50 像素的 Reasonable 子集上，PCA 卷积特征分类器达到了最佳性能 33.49%，相较于基准方法 ACF，提高了 17.87%，比最好的 ACF-SDt 高出 3.95%。图 5-10(b)为所有数据的检测结果。PCA 卷积特征分类器达到了性能 75.65%，相较于基准方法 ACF 提高了 6.28%，但是稍差于最好的 MT-DPM+context。MT-DPM+context 不仅训练了两个不同尺度的分类器而且使用了行人与其它目标关系的约束，所用的信息量更多，也就更适合于 Caltech 这个尺度变化剧烈、遮挡严重的数据集。

图 5-11 展示了 PCA 卷积特征分类器的检测结果，第一行是在 INRIA 上的检测结果，后两行是在 Caltech 上的检测结果。INRIA 数据集中的行人清晰而比较离散，即使有一定的遮挡，也可以有很高的检测率。Caltech 中行人容易受到阴影、复杂背景和尺度等方面的影响，但从图 5-11 可以看出 PCA 卷积特征分类器

图 5-11 行人检测结果

的检测结果还是非常优秀的。

5.4 本章小结

本章介绍了采用对候选区域进行精细分类的新框架在全监督行人检测中的应用。候选区域通过弱分类器获得，在行人这类特殊的目标上，弱分类器比通用候选区域提取方式召回率更高、定位更加精确。精细分类时，本文在聚合通道特征上，提出了基于 PCA 滤波器的卷积网络特征，并使用级联决策树这种 Boost 分

类器进行精细分类。由于卷积核是通过 PCA 学习得到的，所以网络在不进行反向传播的情况下增加了聚合通道特征的判别能力，而且将聚合通道特征进行正交投影更好地配合了 Boost 分类器的特性。

第 6 章 基于渐进优化的自学习行人检测

全监督行人检测中，需要大量的标注样本，工作量庞大。对于特定场景(Scene Specific)的监控视频，由于其可以借助的信息比较多样，可以使用迁移学习和半监督学习[130,131]的方式，但这两种方式都还需要一定量的样本标注，而且标注的样本空间会很大程度限制分类器的学习空间。随着候选区域算法以及若监督分类器算法的不断成熟，通过设计自学习的行人检测器，让其自动的发现行人目标，并迭代更新行人检测器[17]，使学习到的分类器自适应能力更强。本章首先详细介绍自学习行人检测器的设计以及实现，然后通过实验验证自学习检测器在不做任何样本标注的情况下也可以达到有监督行人检测器相近的性能。

6.1 自学习行人检测器

针对特定监控场景的自学习行人检测器，除了必须输入监控场景视频之外，还需要使用一些没有行人的反例图像。这些反例图像的获取成本相比于标定行人样本要低得多，而且反例图像增强了反例样本空间的多样性，有助于促进检测器的学习。由于未使用标注样本，所以自学习行人检测器需要包含一个行人样本发

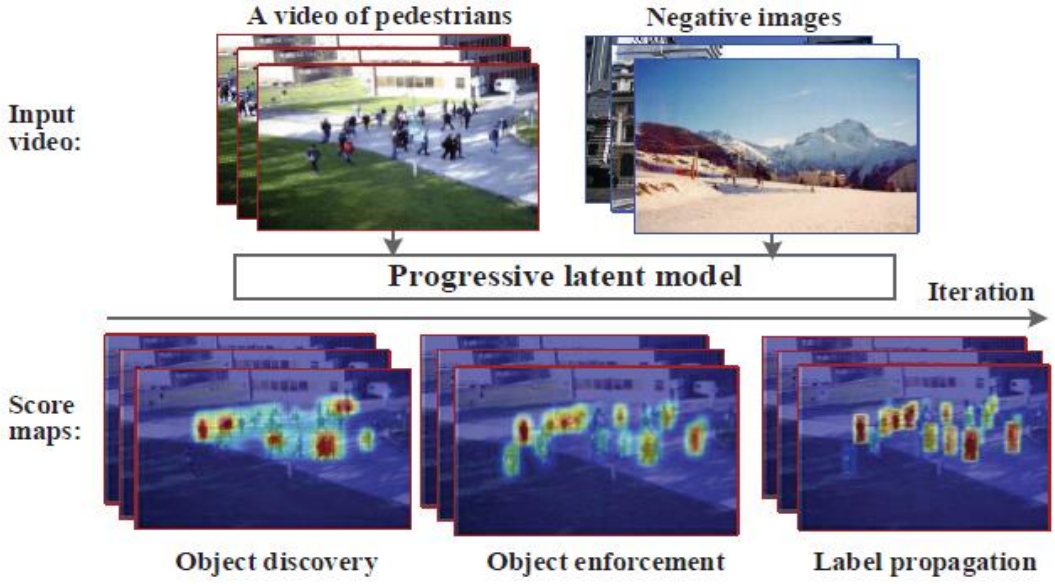


图 6-1 自学习行人检测器的三个阶段

现的阶段；发现的样本并不能保证其一定属于行人样本，所以使用一些场景信息对样本进行增强，取出置信度大的样本当做正例样本；为了扩大检测器的多样性，需要采用标签传播的方式获取更加多样的行人样本。这三个步骤可以迭代进行，学习样本位置 h 和分类器参数 β。由于 h 为隐藏变量，和 β 相互依赖，所以称这个迭代优化模型为渐进隐模型(Progressive Latent Model，PLM)。

定义 $x \in \mathcal{X}$ 为样本空间中的一副图像，样本空间 $\mathcal{X}$ 由特定场景监控视频和反例图像构成； $y \in \mathcal{Y}$， $\mathcal{Y}=\{0,1\}$ 为图像的标注信息。 $y=1$ 表示这一幅图像为正例图像，包含至少一个行人目标，反之 $y=0$ 表示这一幅图像为反例图像，没有行人目标。渐进隐模型的目标函数可以表示为：

$$\begin{aligned}\{h^*,\beta^*\} &= \arg\min_{\beta,h} \mathcal{F}_{(\mathcal{X},\mathcal{Y})}(\beta,h) \\ &= \arg\min_{\beta,h} \mathcal{F}_l(\beta,h) - \lambda\mathcal{F}_s(\beta) + \gamma\mathcal{F}_g(\beta,h),\end{aligned} \tag{6-1}$$

其中 $\mathcal{F}_l(\beta,h)$ 表示行人发现过程中分类误差，越小越好；$\mathcal{F}_s(\beta)$ 表示行人增强时的置信度，越大越好； $\mathcal{F}_g(\beta,h)$ 表示标签传播代价，越小越好， λ 和 γ 是正则项权重。

本小节先详细介绍行人发现，行人增强和标签传播的建模，然后介绍自学习行人检测器的实现。

6.1.1 行人发现

由于没有任何的样本标注，并且不能保证视频序列里的每一帧都包含有行人目标，因此行人发现阶段需要判定哪些视频帧是正例图像，正例图像中哪个位置的窗口为行人目标，学习一个分类器使这个目标的得分最大化。这个优化问题的目标方程可以表示为：

$$\{\mathrm{y}^*,h^*,\beta^*\} = \arg\max_{y\in Y,h\in H,\beta} \beta^T \cdot v(x,y,h), \tag{6-2}$$

其中 $v(x,y,h)$ 表示在视频帧 x 上给定某个类标 y 和某个位置 h 的特征向量。为了降低计算复杂度，隐变量 h 的搜索空间 $\mathcal{H}=\{\mathcal{H}_i\}_{i=1}^N$ 是由行人目标候选区域构成的。

图 6-2 行人发现示意图

对于视频帧，$\mathcal{H}_i$ 表示第 i 帧的目标候选区域集合，对于反例图像，可以直接随机选择一些区域作为反例样本。

通过求解公式(6-2)，会使正例视频帧得到很高的分类得分而抑制反例图像的分类得分。采用隐变量支持向量机(Latent SVM)模型进行求解，并学习到分类器 β，使得分类误差最小

$$\min_{\beta,h} \mathcal{F}_l(\beta,h) = \min_{\beta,h} \frac{1}{2} \| \beta \|^2 + \mathcal{C} \sum_{i=1}^{N} l(\beta, x_i, y_i, h), \tag{6-3}$$

第一项表示最大间隔(max-margin)，第二项为分类器损失，$\mathcal{C}$ 为正则项权重。分类器损失可以表示为：

$$\begin{aligned} l(\beta, x_i, y_i, h) = & \max_{y,h} (\beta^T \cdot v(x_i, y, h) + \Delta(y_i, y)) \\ & - \max_{h} \beta^T \cdot v(x_i, y_i, h) \end{aligned} \tag{6-4}$$

当正确分类，即 $y = y_i$ 时，$\Delta(y_i, y) = 0$，否则 $\Delta(y_i, y) = 1$。该损失函数能够同时选择分类得分大的正例样本的位置 h，并通过分类器 β 使其与其它样本分类距离最大化。

行人发现阶段示意图如图 6-2 所示，该图展示的是执行一次目标发现优化得到的分类器得分的热度图。可以看出采用一轮隐变量支持向量机求解的结果并不

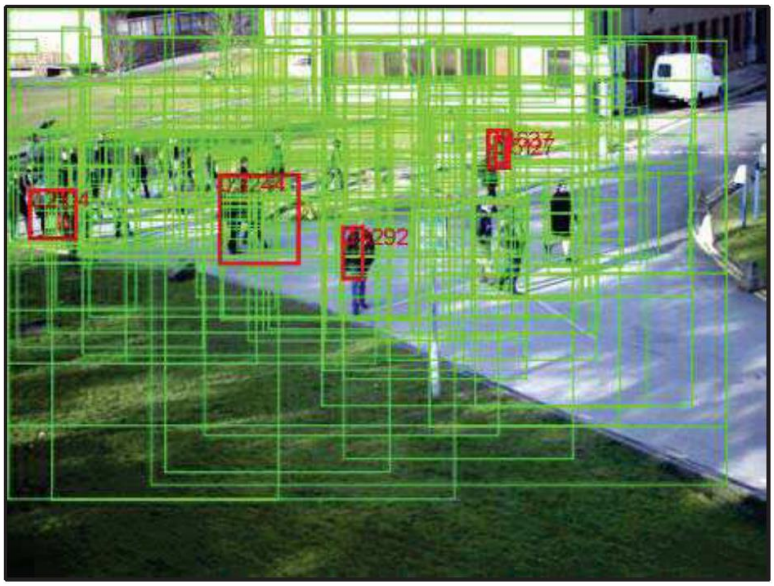


图 6-3 行人增强示意图

理想，分类器得分热度图上基本能够覆盖行人所在的区域，可以判断该图像为正例图像，但并不能很好地获得行人所在的具体位置以及窗口的大小。

6.1.2 行人增强

隐变量支持向量机在学习过程中，更倾向于优化图像级别的分类，一旦存在某些位置 h 使公式(6-3)最小化，优化就会终止[132]。如果位置 h 不能很好地被优化，则很难获得正确的正例行人样本，这一现象在图 6-2(b)中也有所展示。由于隐变量支持向量机的非凸性，优化目标极易陷入局部最优，学习得到的是正例样本的局部区域而不是整个样本。

要使优化跳出这种局部最优，可以通过添加场景信息对其进行约束。类似于 Fast-RCNN[53]中使用与正例样本有一定重叠的区域作为难反例。通过增加难反例样本与正例样本的距离来获得更好的分类器，行人增强阶段也最大化候选区域与其邻域区域之间的距离来进一步优化行人分类器。优化的目标函数为：

$$\max_{\beta} \mathcal{F}_s(\beta) = \sum_{i=1}^{N} \sum_{\substack{h \in \mathcal{H}_i \\ h' \in \Omega_{\mathcal{H}_i,h}}} \| \beta^T \cdot (v(x_i,h) - v(x_i,h')) \|^2 \tag{6-5}$$

其中 $\mathcal{H}_i$ 为前文定义的第 i 帧视频帧获得获选区域，$\Omega_{\mathcal{H}_i,h}$ 表示给定候选区域位置 h 的在 $\mathcal{H}_i$ 空间中的邻域。和 Fast-RCNN 设置一样，将与 h 交叠比(IoU, Intersection

over Union)在 0 到 0.25 之间的作为其邻域窗口，也即难反例样本。这些邻域窗口往往是目标的局部区域，或者包含有目标的更大的窗口区域。

公式(6-5)的优化是在每个给定的候选区域位置 h 的情况下计算的，这是一个凸优化问题，得到的分类器模型 β 可以被用来挑选正例样本。行人增强示意图如图 6-3 所示，可以从候选区域空间 $\mathcal{H}_i$ 中获得一些孤立的高得分样本。

6.1.3 标签传播

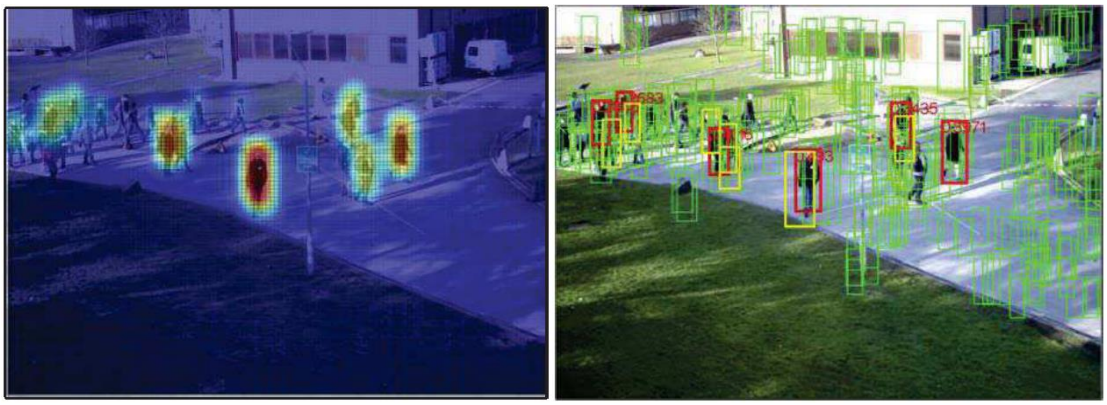
经过分类器学习以及分类器增强，虽然可以得到一些孤立的高得分样本进一步进行训练，但只是迭代这两步会产生一些弊端：1) 即使选择的样本是完全正确的，由于其选择的样本数量少且多样性差，分类器的泛化能力变差；2) 如果存在一些选择错误的样本，会使分类器朝着错误的方向越走越远。在第一次迭代时不仅样本少，而且容易出现错误样本，如图 6-3 所示，所以需要采用增量学习来增加样本多样性，同时使正例样本包中真正的正例样本含量增加。自学习行人检测器算法中采用标签传播作为增量学习算法。

假设通过行人发现和行人增强之后获得了 l 个被标记为正例的样本，选取 $u = l \times (r - 1.0)$ 个未被标记的样本。$r > 1.0$，取决于监控场景中的行人密度，当场景中预计行人密度比较高时取比较大的 r 值，以获得更多的样本，反之则取较小的 r 值，意味着被标记的 l 个样本对分类器有更强的影响力。将有标记和未标记的样本记做 $\{h_i\}_{i=1}^{l}$ 和 $\{h_j\}_{j=l+1}^{l+u}$，这些样本可以构建一个 kNN 图，图的顶点就是样本点。当且仅当 h_i 在 h_j 的 kNN 图中或者 h_j 在 h_i 的 kNN 图中[133]时，h_i 和 h_j 有边连接。

基于图的标签传播过程定义为：

$$g(\beta, h_j) = \frac{\sum_{i=1}^{l} w_{ji} g(\beta, h_i)}{\sum_{i=1}^{l} w_{ji}} \tag{6-6}$$

其中 w_{ji} 表示样本 h_j 和 h_i 之间的连边权重。标签传播可以通过凸优化进行求解[133]，其目标函数为：



(a) (b)

图 6-4 渐进优化结果

$$\max_{g(\beta,h)} \mathcal{F}_g(\beta,h) = \min_{g(\beta,h)} \sum_{i=l}^{l} \sum_{j=l}^{l+u} w_{ij}(g(\beta,h_i) - g(\beta,h_j))^2$$
$$s.t. \quad g(\beta,h_i) = y_i, i = 1,...,l, \tag{6-7}$$

目标函数最小化被标记和未被标记样本标签之间的距离，同时受到已标记样本标签已知的约束。通过标签传播算法可以选择出更多的样本以增加分类器的多样性和容错率。

6.1.4 自学习行人检测器实现

模型求解：由于公式(6-1)中行人增强目标函数 $\mathcal{F}_s(\beta)$ 和标签传播目标函数 $\mathcal{F}_g(\beta,h)$ 的优化需要依赖于行人发现目标函数 $\mathcal{F}_l(\beta,h)$ 的优化结果，所以采用渐进求解算法，交替优化这三个目标函数。

在公式(6-3)中，$\mathcal{F}_l(\beta,h)$ 可以写成 $A(x)-B(x)$，那么自学习行人检测器的整体优化目标函数 $\mathcal{F}$ 可以被写成 $A(x)-B(x)+C(x)-D(x)$ 的形式。这种形式的目标函数可以采用 Concave-Convex(CCCP)求解[132]。CCCP 算法第一步，通过优化 $\mathcal{F}_l(\beta,h)$ 在视频帧中发现潜在的行人目标，初始化隐模型。CCCP 算法的第二步，通过优化 $\gamma\mathcal{F}_g - \lambda\mathcal{F}_s$ 进行行人目标增强和标签传播，来选择更准确多样性更强的行人样本。通过这两步迭代优化，CCCP 可以使整体目标函数收敛到局部最小值或者鞍点上。

渐进优化的结果如图 6-4 所示，该图展示的是 CCCP 算法执行到第五轮迭代

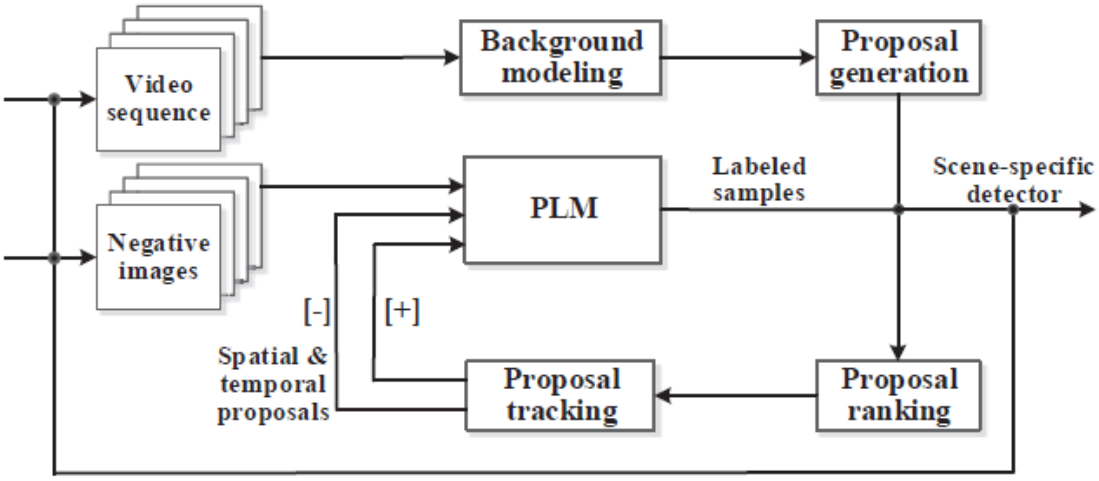


图 6-5 自学习检测器流程图

后的结果。在同一帧图像上，相比于图 6-2 和图 6-3 中第一次迭代的结果，图 6-4(a)中的分类得分热度图从成片的行人区域或者行人局部区域逐步聚集到一些完整孤立的行人区域上，这也导致图 6-4(b)中通过行人目标增强和标签传播之后获得的样本更加准确，而且行人的样本数量增加使多样性得到充分保证。验证了分步迭代的 CCCP 算法能够持续地降低样本的误差，保证自学习分类器的稳定性。

实现细节：图 6-5 展示了自学习行人检测器的实现流程。Proposal generation 通过使用目标的运动以及外观信息，可以初步获得候选目标区域。Proposal ranking 将得分高的候选区域选作正例样本，得分低的样本作为反例样本。Proposal tracking 利用视频帧的信息进一步加强候选区域的正确性。渐进隐模型 PLM 根据上一轮获得正反例样本训练 DPM(Deformable Part-based Model)模型[23,52]作为行人检测器。

对于给定的特定场景监控视频，由于其视角焦距固定，可以通过背景建模获得极强的先验信息。但背景建模只能获得动目标大致区域，并不能获得样本窗口。为了解决这个问题，在候选区域获取的初始化阶段，使用了本文提出的目标预定位方法获得一些窗口。假设一个窗口中的完整边缘含量越多的话，越有可能包含一个感兴趣的目标物体。完整边缘的数量可以直接用来衡量窗口包含行人的置信度。在第一轮迭代时，候选区域通过背景建模动目标置信度和预定位区域的置信度进行综合排序。这些被选择出来的候选区域作为样本训练一个检测器。到第二轮以后，除了背景建模和预定位的置信度之外，还可以得到检测器得分。从第二

帧开始就可以利用视频中的跟踪信息进一步确定正例样本的正确性。采用 KLT 跟踪算法[134]从t到$t+\tau$ 帧中来跟踪和收集候选区域，这里τ根据经验设定为 10。

假设 $f_\beta(h)$、 $f_m(h)$和 $f_o(g)$分别表示检测器得分、运动得分和 EdgeBoxes 分数，候选区域排序的得分定义为：

$$f(h)=\alpha^T\cdot(f_\beta(h),f_m(h),f_o(g)) \tag{6-8}$$

其中α^T为权重向量。运动得分 $f_m(h)$为候选区域中各个像素的运动得分平均值，可以直接通过背景建模图像快速计算得到。区域预定位得分 $f_o(g)$在通过[34]获得候选区域时就同时计算得到。检测器得分 $f_\beta(h)$在第一轮迭代时无法获得，只有到第二轮之后才能通过训练的检测器计算出各个窗口的得分。通过训练得到检测器之后，为了增强行人窗口的定位准确性，以预定位得到的各个候选区域的中心为中心，采用多尺度滑窗获得更加细密的候选区域。

权重向量α^T初始化为$(0,0.5,0.5)$表示初始时不考虑检测器得分。从第二轮开始，采用零空间回归(zero-space regression method)[135]进行更新。零空间回归在最小化所有候选区域的回归误差的同时，最大化新的超平面与原始超平面之间的距离。零空间回归方法不仅可以在输入的候选区域中选择出概率较大的样本，而且可以增强候选区域排序的自适应性。

错误率分析：渐进隐模型(PLM)中引入了标签传播来获得更多的训练样本，以此来更新行人检测器。标签传播中，最主要的问题是如何确定其传播范围使行人检测器在不发散的同时还降低检测错误率。通过公式(6-1)，更大的权重γ意味着产生更多的正例训练样本，同时也增加了分类错误率ξ，反之亦然。通过标签传播选择出的未标记样本的个数u可以设置为权重γ的函数，即$u(\gamma)$。γ的选择不仅需要保证新选择出的正例训练样本得到的分类器的错误率比原始正例样本得到的分类器更好，即检测的错误率维持稳定或者单调递减。同时也希望γ的选择能尽可能得大，让更多的训练样本被选择出来，以增强多样性。因此渐进隐模型中采用优化的方式确定γ，其目标方程为：

$$\begin{aligned} &\max_{\gamma,\beta,y_j} \gamma \\ &s.t. \quad \xi_{u(\gamma)} \le \xi_l \\ &\approx \frac{1}{l+u(\gamma)} \sum_{j=1}^{l+u(\lambda)} (f_\beta(h_j) - \tilde{y}_j) \le \frac{1}{l} \sum_{i=1}^{l} (f_{\tilde{\beta}}(h_i) - \tilde{y}_i), \end{aligned} \tag{6-9}$$

其中l和$u(\gamma)$为标签传播之前已被标记的样本个数和通过标签传播标记的样本个数。这个优化目标在最大化γ的同时受到了错误率不升高的约束，可以采用线搜索[136]进行求解，搜索空间被限制在[0,1.0]这个闭区间内，搜索的步长设置为 0.1。在每个搜索步骤中，由于增加了新的未标记样本，所以需要对$f_\beta(h_j)$更新为$f_{\tilde{\beta}}(h_j)$。获得了更新之前和之后的分类器，就可以计算出错误率。

6.2 实验结果及分析

实验中，首先比较了行人增强和标签传播在自学习检测器中的作用，以及其作为正则项时的权重参数的影响，然后与有监督以及弱监督的行人检测方法进行了比较与分析。

6.2.1 实验配置

数据集：除了已有的特定场景监控视频的四个公开数据集，还采集了一个更加有挑战性的 24 小时监控数据集。

PETS2009[137]为校园场景下的视频序列，其分辨率为 720x576。背景信息相对简单，行人目标分辨率也较高，但存在多个目标且有严重的遮挡。

Towncenter[138]为市中心的视频序列，其分辨率为 1920x1080。除行人目标多以外，其挑战有场景大导致的行人分辨率较小，而且略微俯视的视角导致了行人的变形。

PNN-Parking-Lot2/Pizza[11]有两个视频序列，分辨率为 1920x1080。其行人的数量不是太多，但由于复杂路线以及多组行人人群，产生了大量的姿态变化和遮挡，为检测带来了挑战。

CUHK Square[69]为一个十字路口的监控室视频，长达 60 分钟，分辨率为 704x576。该数据集的场景相对比较复杂，除了复杂的背景之外，运动的车、接近 45 度的俯视视角、行人样本分辨率低以及行人出现不连续，都为行人检测产

生了阻碍。

24Hours 数据集顾名思义是一个包含了 24 小时长度的监控视频，其分辨率为 704x576。由于视频帧含量极大，为了降低计算复杂度，均匀抽取了 6000 帧作为训练数据，2600 帧作为测试数据。这个 24 小时的监控视频更加接近真实使用环境，存在很强的照明变化、时而密集时而稀少的行人片段、以及其它运动物体的干扰。

评测标准：在 PETS2009、Towncenter、PNN-Parking-Lot2/Pizza 和 24Hours 这四个数据集上，采用 Precision-Recall 曲线来进行评估，曲线越靠近右上角检测性能越好。在 CUHK 数据集上，与提出该数据集的作者采用的评价指标一致，采用 Recall-FPPI 曲线进行评价[69]，在每幅图像的误警率(FPPI, False Poitive Per Image)相同时，召回率越高则检测性能越好。

6.2.2 自学习检测器分析与参数选择

目标增强：由于在目标发现过程中的优化目标函数(6-3)为非凸函数，容易陷入局部最小或者鞍点，如图 6-2 所示也确实出现了这种现象。通过目标增强之后，可以一定程度优化检测器的性能。如图 6-6 所示，在数据集 PETS2009 中使用目标增强后，当召回率在 0.7 时，检测精度提升了 10%以上，表明目标增强确实有助于抑制陷入行人目标局部区域、成片行人区域的样本以及被误检的样本(False Positive)。

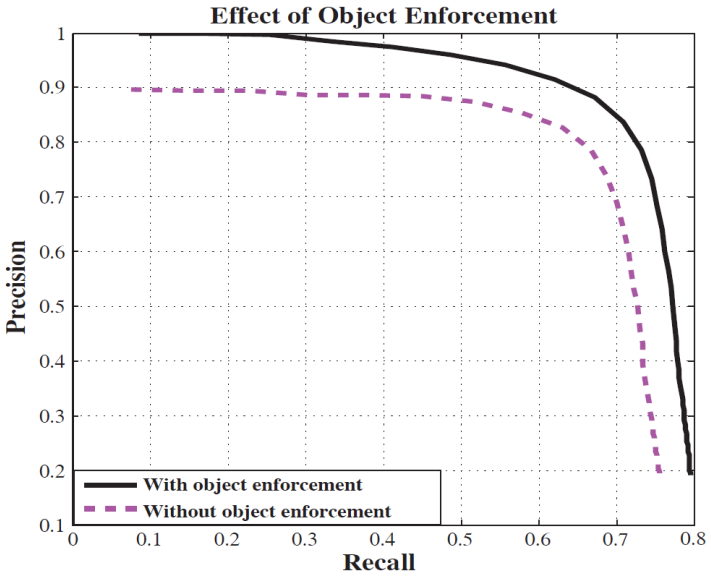


图 6-6 行人增强对自学习行人检测器的影响

标签传播：在自学习检测器中，标签传播目的是在增量式增加行人样本的同时降低错误率。在实验中，同时评测了这两个方面。

在 PETS2009 数据集上，标签传播对于分类器多样性的改进如图 6-7 所示。每次迭代都可以观测到明显的性能提升，召回率的提升说明分类器的多样性得到了增强，精度的提升说明了分类性能得到了增强。经过十多次迭代之后，就没有更多的未被标记的样本被选择为正例训练样本，检测器的性能也趋于稳定，接近最终的检测结果。

错误率的评测如图 6-8 所示，同时展示了所有五个数据集六段视频序列中的变化。可以看到在所有视频序列上错误率都趋于下降。当迭代到十轮之后，错误率也趋于收敛，从另一个角度证明了不再增加新的样本。

权重向量α^T：图 6-9 展示了权重向量α^T在 PETS2009 数据集上随着迭代过程中的变化。由于自学习行人检测器针对的是场景固定的检测视频，所以通过背景建模获得的运动信息非常可靠，权重一直很稳定，如蓝色虚线所示。随着迭代的进行，分类器得分的权重在不停的上升，说明分类器的分类性能越来越好，在选择样本的时候重要性也依次增加。而 EdgeBoxes 得分随着迭代的进行，迅速衰减到 0，意味着 Edgeboxes 的得分对于筛选候选区域有效，但对行人分类却没有那么强的判别性。

参数选择：表 6-1 展示了在所有五个数据集上γ的选择。由于 Towncenter 数据集中光照变化不明显以及运动干扰比较少，而且行人特别稀疏导致每帧获得的训练样本特别少，所以可以选择很大的γ，达到了 0.70。CUHK 数据集和 24Hours 这两个数据集的γ很小，为 0.30，说明这两个数据集中受到其它运动以及光照的干扰比较大。

6.2.3 自学习检测器检测结果

除 24Hours 数据集中已经分类了训练集和测试集之外，其它数据集中使用一半视频帧用于训练，另一半视频帧用于测试。为了评估自学习行人检测器的性能，采用了以下有监督学习、迁移学习和弱监督学习方法进行比较。所有性能都是基于以下方法提供的代码进行重新实现获得。

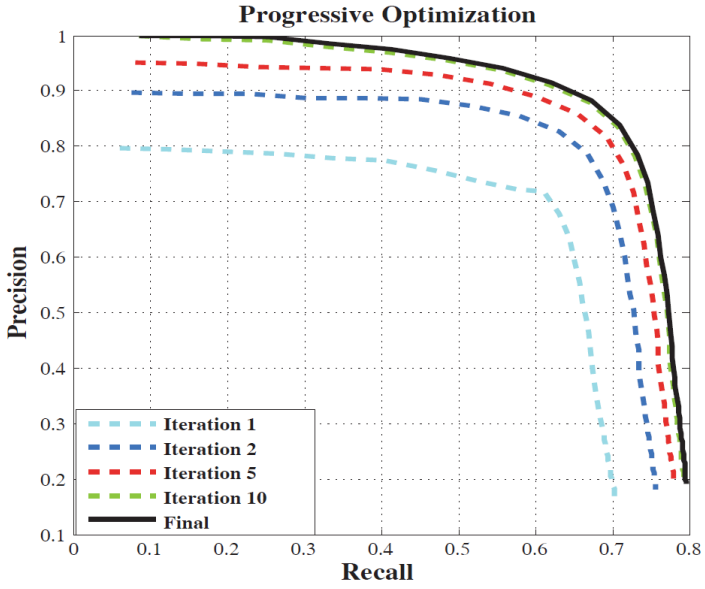


图 6-7 标签传播对自学习行人检测器的影响

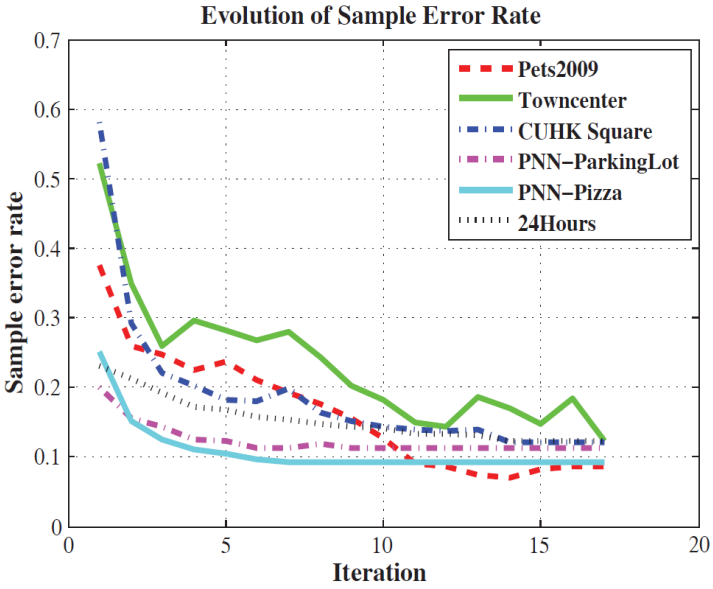


图 6-8 自学习行人检测器的错误率

Dataset	PETS	Towncenter	PNN	CUHK	24Hours
γ	0.50	0.70	0.60	0.30	0.30

表 6-1 自学习行人检测器正则项参数选择

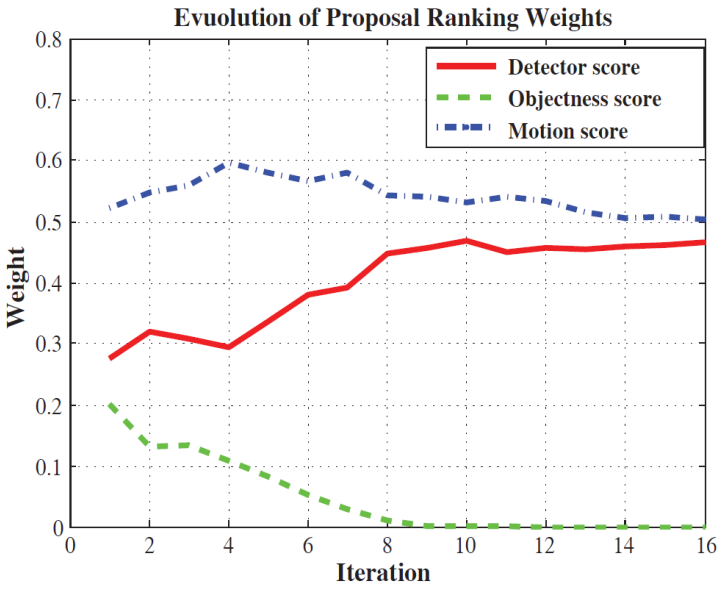


图 6-9 分类器、运动和 Edgeboxes 对候选区域排序的影响

Offline-DPM[23,52]：离线 DPM 模型指直接使用 PASCAL VOC 数据上训练的人这类目标的检测器进行检测。

Supervised-DPM[23,52]：仍然使用 DPM 训练检测器，不同的是直接使用上述五个数据集中标准的行人样本进行有监督的

训练，同时从反例图像中挖掘难反例样本以获得更好的检测性能。

Supervised-SLSV[130]：在特定场景行人检测中取得了极佳性能的方法。其利用一些与特定场景中行人外观相似的合成样本进行分类器训练。但该方法未公开源代码，因此只给出了其评测的 Towncener 数据集上的性能。

Transfer-DPM[11]：该方法通过 Offline-DPM 在 PASCAL VOC 数据上学习到人的检测器，然后通过迁移学习(Transfer Learning)将 Offline-DPM 迁移到新的数据集上。

Transfer-SSPD[69]：使用迁移学习方式在场景特定的行人检测中取得极佳检测性能的方法。

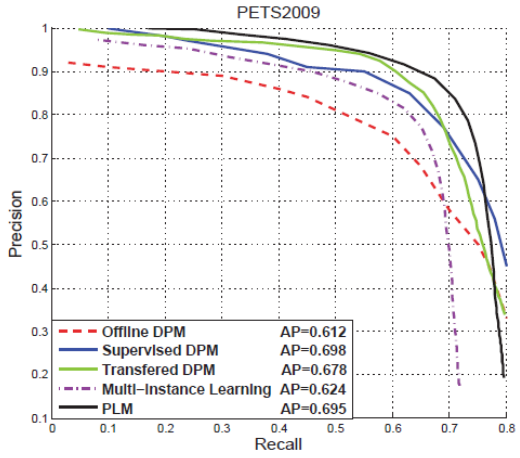
Weakly-MIL[12]：在弱监督学习中取得极大成功的多实例学习(Multiple Instance Learning)，在获得标记的正例样本之后，仍然采用 DPM 训练行人检测器。

PETS2009、Towncener、PNN−Parking−Lot2、PNN−Pizza、CUHK Square 以及 24Hours 数据集上的性能分别如图 6-10(a)到图 6-10(f) 所示。从图中可以看出，

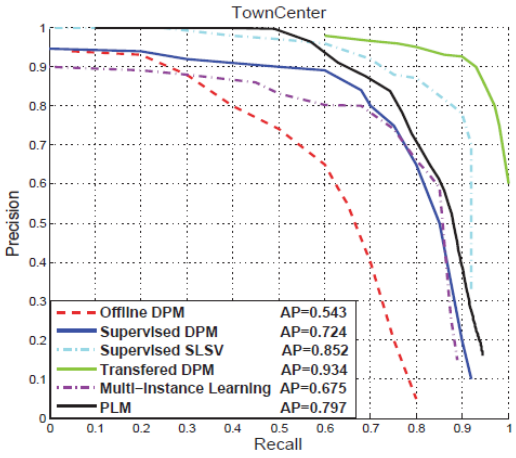
在 Precision-Recall 曲线以及 Recall-FPPI 曲线上，自学习检测器的性能都显著优于 Offline-DPM。由于行人只关注站立的人，而 PASCAL VOC 数据集中含有各种姿态的人，因此其在特定场景行人检测上性能不高，也说明了数据集差异为检测带来了极大的挑战，自学习行人检测就更加有意义。自学习检测器也显著优于弱监督学习的 Weakly-MIL 方法。

在 PETS2009 和 PNNParking-Lot2 数据集上，自学习检测器获得了最佳的性能，如图 6-10(a)和图 6-10(c)所示。在 Towncener 数据集上，自学习检测器略低于使用合成样本的 Supervised-SLSV 方法，也低于迁移学习，但仍然高于弱监督的方法，如图 6-10(b)所示。自学习检测器在这个数据集上未达到较好性能的原因可能是因为该数据集上行人密度太稀疏，导致了样本选取上的困难。值得强调的是，有监督学习和迁移学习都需要标注的样本，而自学习检测器并不需要任何的样本标注。在 PNN-Pizza 数据集上和 CUHK Square 数据集上，自学习检测器显著高于迁移学习的方法，甚至与有监督的检测器 Supervised-DPM 性能可比，如图 6-10(d)和图 6-10(e)所示。在 24Hours 数据集上，自学习行人检测器的平均检测精度(AP)为 0.702，高于其它所有的方法，如图 6-10(f)所示。比迁移学习高出 6 个百分点，说明迁移学习在复杂光照环境下确实受到比较大的干扰，而自学习行人检测器可以迭代地选择不同光照条件下的样本，使分类具有更高的适应性。虽然全监督的学习方式也可以选择到不同光照条件下的样本，但自学习行人检测器的检测性能比全监督的方式更好。

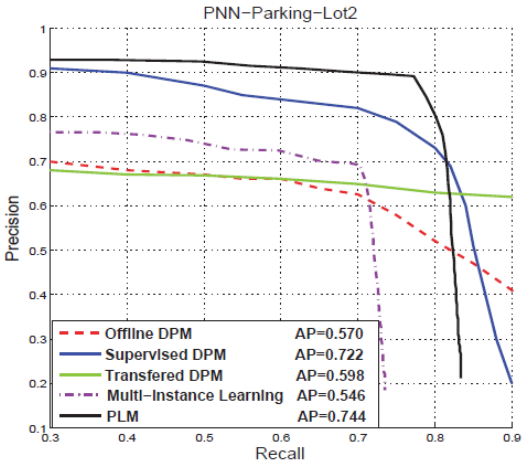
自学习检测器的迭代展示和检测结果如图 6-11 所示，其中第一列到第三列分别为第一次迭代、第五次迭代和第十次迭代的结果，第四列为经过第十次迭代后获得的样本。通过前三列的比较，可以看到正例样本越来越多，越来越精确，也从成片的区域迭代为离散的行人区域。在 PNN-Parking-Lots2 这个场景相对干净的数据中，所有的正例样本都被准确地选择出来。在 PETS2009 以及 PNN-Pizaa 这两个拥挤的数据集上，虽然未能正确标记出所有的行人样本，但所选择出来的样本位置也很准确。在 Towncenter 和 CUHK square 数据集上，虽然存在其它运动目标的干扰，但仍然能选择出准确的样本，说明自学习行人检测器对噪声具有鲁棒性。24Hours 数据集则因其时间从清晨到午夜，行人时多时少，分辨率低等



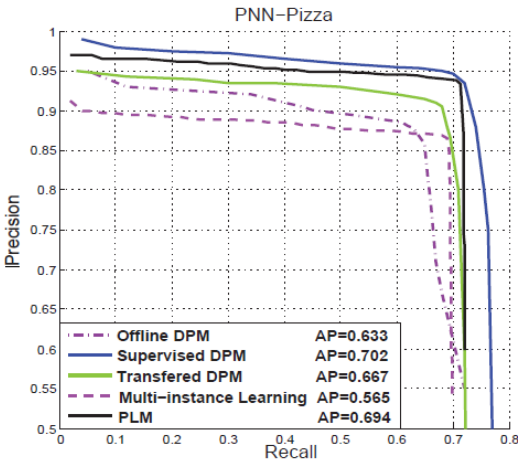
(a)



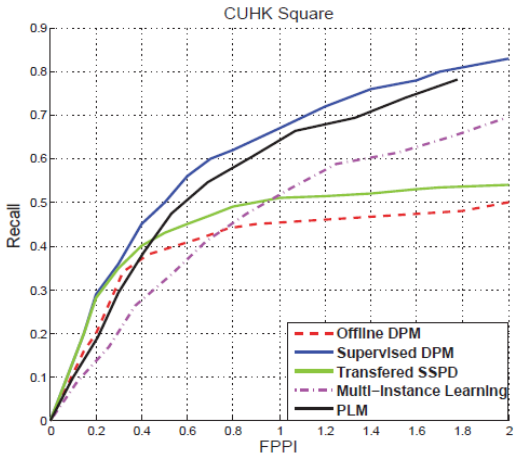
(b)



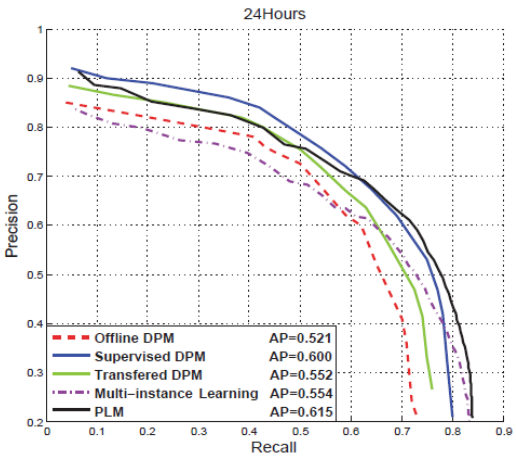
(c)



(d)



(e)



(f)

图 6-10 自学习行人检测在公开数据集上的评测

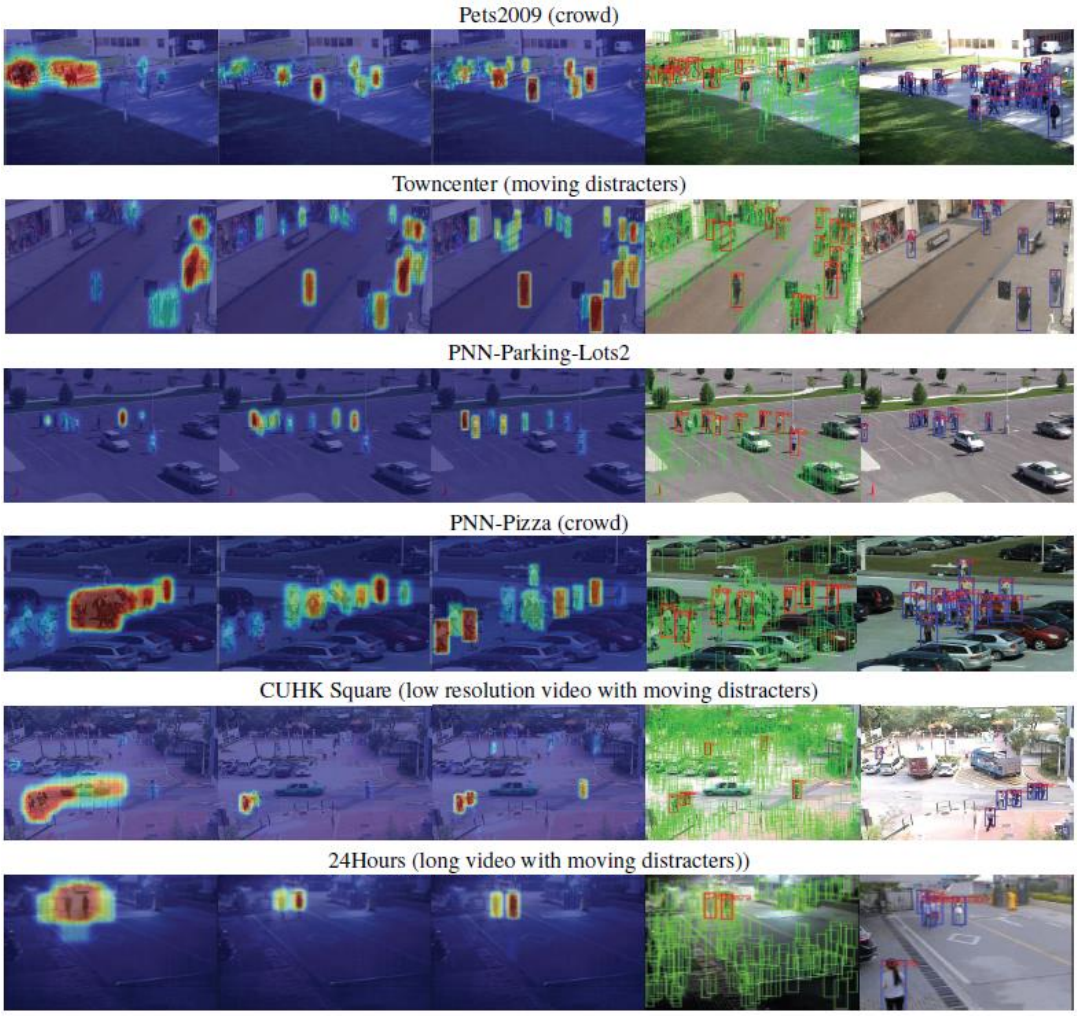


图 6-11 自学习行人检测器的迭代展示及检测结果

(a) 迁移学习　　(b) 自学习分类器

图 6-12 自学习检测器在 24Hours 数据集上的检测结果

原因最具有挑战性，图中展示在行人较少时能够完全正确的选择行人样本。

图 6-11 的最后一列为测试结果，展示了自学习分类器在特定场景监控视频中的检测结果。从结果可以看出，其对遮挡、低分辨率、视角变化和姿态变化都具有很强的鲁棒性。特别地，图 6-12 展示了 24Hours 数据集上白天和傍晚两种情况下，自学习行人检测器与基于迁移学习的检测器的检测结果比较。可以看出

左侧自学习检测器能够适应光照的变化，而右侧的迁移学习方法会出现错检测。

6.3 本章小结

针对全监督行人检测器的训练需要消耗大量人力进行样本的标注的问题，本章提出了在尽可能不使用样本标注的情况下，自动地在特定场景的监控视频中学习行人检测器的自学习算法。相比于迁移学习和半监督学习，自学习算法能够自动更新获得的样本以增加分类器的适应性。自适应分类器使用渐进优化模型，在初始化目标候选区域后，同时从空间域和时间域约束优化候选区域并采用标签传播增加多样性，迭代地训练以目标位置为隐变量的隐模型。实验表明，本文所研究的自学习行人检测器能够达到可比于全监督检测器的性能，为自学习相机的推进打下了坚实的基础。

第 7 章 总结与展望

行人检测可应用于智能视频监控、车辆辅助驾驶以及智能机器人等方面，也可为图像视频检索、目标跟踪和图像分类等其它计算机视觉相关领域提供重要信息，具有重要的研究价值和实际应用价值。本文主要研究行人检测中底层和中层特征表示，以及其在全监督和自学习行人检测中的应用。描述精确的底层特征能够用来获得行人候选区域，在全监督行人检测中可以降低检测时间，在自学习分类器中可以降低样本搜索空间。表示能力强的中层特征能够增强检测模型的判别性，提高全监督行人检测的性能。

7.1 本文工作总结

对于行人底层特征提取，本文提出了基于侧输出残差网络（SRN）的对称性信息和边缘信息的提取方法，并将其应用于候选区域排序。对于行人中层特征提取，本文提出基于 PCA 卷积特征（PCA-CCF）并将其用于全监督行人检测。虽然候选区域提取方法在全监督行人检测中比直接采用弱分类器获得的区域效果差，但其可以被用于初始化自学习分类器的样本搜索空间，使自学习分类器更快更准确地收敛。本文主要研究成果总结如下：

(1) 侧输出残差网络使用简单有效的残差单元，拟合残差单元的输入与真实值之间的残差。利用卷积神经网络的多尺度结构，依次自深到浅地堆叠残差单元，使残差依次变小。该方法采用了经典的由粗到精的思路，在给定一个初始预测输出后，自适应地选择合适尺度的响应对初始输出进行不断的优化，以降低错误率。实验表明，侧输出残差网络简单有效，能够极大地提高基准方法的对称性或边缘提取的准确性，而且能够很容易扩展成多分支结构，提取输入图像的边缘和对称性信息。

(2) 结合超像素合并方式和置信度方式的目标候选区域提取方法的优点，提出了基于贝叶斯得分重排序的候选区域提取方法。首先使用相似性自适应搜索这种超像素合并方式获得冗余的区域，然后使用提取的对称性信息和边缘信息计算

对称性得分和边缘得分并将其在贝叶斯框架下统一为概率模式，衡量给定窗口包含感兴趣目标的置信度。并按照置信度对冗余区域进行重排序，以达到在用更少候选区域的同时保持其召回率和定位精度。

(3) 使用 PCA 基的正交特性，本文提出了 PCA 卷积特征（PCA-CCF）。PCA-CCF 通过张量运算增加了聚合通道特征的表达能力，而且将聚合通道特征各个通道进行去相关，以适应级联决策树分类器。由于增加了特征提取的计算复杂度，因此采用由粗到精的方式，先用弱分类器获得数个候选区域，再利用 PCA-CCF 对这些区域进行精细分类。实验表明，行人检测的性能相比于聚合通道特征有显著的提高。

(4) 在自学习行人检测中，本文使用目标候选区域学习检测器；利用候选区域置信度、检测器得分、运动得分对样本进行增强；使用标签传播选择更多的正例样本和难反例样本。渐进迭代这些步骤，不断优化检测器的检测能力。在特定静态视频场景中，本论文研究的方法可以实现不使用任何样本标注训练行人检测器的目的。

7.2 未来工作展望

随着智能监控、无人驾驶以及智能机器人这些应用的驱动，行人检测方数据集的约束越来越小，已经从刚开始使用样本图像，到使用拍摄的照片，再到特定的监控视频，直至复杂的交通场景。行人检测的性能会成为制约其应用的重要因素，而行人的底层和中层特征提取又制约着性能的提升。未来可以从如下几个方面进行后续研究，完善行人检测系统：

(1) 单目标候选区域的提取：现有的候选区域提取都是针对通用目标而言的，对于某类特殊的单目标，通用目标的候选区域会存在非常多的干扰项。设计更加有针对性的单目标底层特征，将每幅图像的单目标候选区域降低到数十个或者数个，将会极大地提升全监督检测的速度以及无监督检测的搜索空间。本文中的候选区域提取方法在每幅图像上要达到可以受收的召回率时仍需要数百个窗口，全监督时不及弱分类器，自学习时可能导致算法发散。

(2) 快速的中层特征提取：中层特征制约着分类器精度，但现有卷积神经网络特征虽然表现能力强，但往往计算复杂度很高。本文的 PCA-CCF 特征只使用了两层前馈结构就极大地增加了特征提取时间。如果中层特征能够在图像上提取一次，每个窗口只需要使用查表法就可以快速获得窗口的特征，成倍增加检测速度，这一想法已在 RoIPooling[53]中得到验证。但 RoIPooling 还仍未达到真正的查表法速度，因为其需要对窗口的特征进行特征对齐。如果中层特征能够极快地提取，也会促进行人检测速度的提升。算法和硬件相辅相成，或许硬件运算速度的加强会直接解决行人检测速度的问题。

(3) 自学习行人检测及其应用：数据量的急速增长为自学习行人检测器提供了前所未有的机会。大量数据样本标注成本高，无论多大的样本库都不能覆盖整个正例样本空间。自学习检测器以其自动发现新样本更新检测器成为解决这个问题的最佳方案之一。目前自学习检测器算法更新样本和检测器需要很长的计算时间，稀疏而大动态的行人视频也给自学习检测器带来挑战。成功的自学习检测可以植入智能监控前端。如果每个监控摄像头都能够学习到自己特定的检测模型，会极大地增强智能监控摄像头的场景适应性。

参考文献

[1] Viola P. A., Jones M. J.. Robust real-time face detection[J]. *International Journal of Computer Vision*, 2004, 57(2): 137–154.

[2] Dalal N., Triggs B.. Histograms of oriented gradients for human detection[C]. *In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, 2005: 886–893.

[3] Everingham M., Van Gool L., Williams C. K. I., et al. The PASCAL Visual Object Classes Challenge 2011 (VOC2011) Results[M]. http://www.pascalnetwork.org/challenges/VOC/voc2011/workshop/index.html

[4] Angelova A., Zhu S.. Efficient object detection and segmentation for fine-grained recognition[C]. *In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, 2013: 811–818.

[5] Krizhevsky A., Sutskever I., Hinton G. E.. Imagenet classification with deep convolutional neural networks[C]. *In: Proceedings of Advances in Neural Information Processing Systems*, 2012: 1106–1114.

[6] Lin T., Maire M., Belongie S. J., et al. Microsoft COCO: common objects in context[C]. *In: Proceedings of European Conference on Computer Vision*, 2014: 740–755.

[7] Cortes C., Vapnik V.. Support-vector networks[J]. *Machine Learning*, 1995, 20 (3): 273–297.

[8] Gelman A., Carlin J. B., Stern H. S., et al. Bayesian data analysis: volume 2[M]. 2014.

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

[10] Liaw A., Wiener M.. Classification and regression by random foreast[J]. *R news*, 2002, 2(3): 18–22.

[11] Shu G., Dehghan A., Shah M.. Improving an object detector and extracting regions using superpixels[C]. *In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, 2013: 3721–3727.

[12] Cinbis R. G., Verbeek J. J., Schmid C.. Weakly supervised object localization with multi-fold multiple instance learning[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017, 39(1): 189–203.

[13] Girshick R. B., Donahue J., Darrell T., et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]. *In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, 2014: 580–587.

[14] Hinton G. E., Salakhutdinov R. R.. Reducing the dimensionality of data with neural networks[J]. *science*, 2006, 313(5786): 504–507.

[15] Chen X., Wei P., Ke W., et al. Pedestrian detection with deep convolutional neural

network[C]. *In: Proceedings of Asian Conference on Computer Vision Workshops*, 2014:354–365.

[16] Ke W., Zhang Y., Wei P., et Al. Pedestrian detection via PCA filters based convolutional channel features[C]. *In: Proceedings of IEEE International International Conference on Acoustics, Speech and Signal Processing*, 2015: 1394–1398.

[17] Ye Q., Zhang T., Ke W., et Al. Self-learning scene-specific pedestrian detectors using a progressive latent model[C]. *In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, 2017: 2057–2066.

[18] Ke W., Zhang T., Chen J., et al. Texture complexity based redundant regions ranking for object proposal[C]. *In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition Workshops*, 2016: 1083–1091.

[19] Ke W., Chen J., Ye Q.. Deep contour and symmetry scored object proposal[J]. *Pattern Recognition Letters (Accepted)*, 2018.

[20] Ke W., Chen J., Jiao J., et al. SRN: side-output residual network for object symmetry detection in the wild[C]. *In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, 2017: 302–310.

[21] Ke W., Chen J., Jiao J., et al. SRN: Side-output residual network for object symmetry detection and Beyond[J]. *CoRR*, 2017.

[22] Liu C., Ke W., Jiao J., et al. RSRN: rich side-output residual network for medial axis detection[C]. *In: Proceedings of IEEE International Conference on Computer Vision Workshops*, 2017: 1739–1743.

[23] Felzenszwalb P. F., Mcallester D. A., Ramanan D.. A discriminatively trained, multiscale, deformable part model[C]. *In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, 2008.

[24] Girshick R. B.. Fast R-CNN[C]. *In: Proceedings of IEEE International Conference on Computer Vision*, 2015: 1440–1448.

[25] Liu W., Anguelov D., Erhan D., et al. SSD: single shot multibox detector[C]. *In: Proceedings of European Conference on Computer Vision*, 2016: 21–37.

[26] Redmon J., Divvala S. K., Girshick R. B., et al. You only look once: Unified, real-time object detection[C]. *In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, 2016: 779–788.

[27] Redmon J., Farhadi A.. YOLO9000: better, faster, stronger[C]. *In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, 2017: 6517–6525.

[28] He K., Zhang X., Ren S., et al. Spatial pyramid pooling in deep convolutional networks for visual recognition[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2015, 37(9): 1904–1916.

[29] Zhang L., Lin L., Liang X., et al. Is faster R–CNN doing well for pedestrian detection?[C].

In: Proceedings of European Conference on Computer Vision, 2016: 443–457.

[30] Hosang J. H., Benenson R., Schiele B.. How good are detection proposals, really?[C]. *In: Proceedings of British Machine Vision Conference*, 2014.

[31] 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.

[32] Van De Sande K. E. A., Uijlings J. R. R., Gevers T., et al. Segmentation as selective search for object recognition[C]. *In: Proceedings of IEEE International Conference on Computer Vision*. 2011: 1879–1886.

[33] Cheng M., Zhang Z., Lin W., et al. BING: binarized normed gradients for objectness estimation at 300fps[C]. *In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, 2014: 3286–3293.

[34] Zitnick C. L., Dollár P.. Edge boxes: Locating object proposals from edges[C]. I*n: Proceedings of European Conference on Computer Vision*, 2014: 391–405.

[35] Szegedy C., Liu W., Jia Y., et al. Going deeper with convolutions[C]. *In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, 2015: 1–9.

[36] Papageorgiou C., Poggio T. A.. A. trainable system for object detection[J]. *International Journal of Computer Vision*, 2000, 38(1): 15–33.

[37] Lowe D. G.. Distinctive image features from scale–invariant keypoints[J]. *International Journal of Computer Vision*, 2004, 60(2): 91–110.

[38] Zhu Q., Yeh M., Cheng K., et al. Fast human detection using a. cascade of histograms of oriented gradients[C]. *In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, 2006: 1491–1498.

[39] Ye Q., Jiao J., Zhang B.. Fast pedestrian detection with multi–scale orientation features and two–stage classifiers[C]. *In: Proceedings of IEEE International Conference on Image Processing*. 2010: 881–884.

[40] Jia H. X., Zhang Y. J.. Fast human detection by boosting histograms of oriented gradients[C]. *In: Proceedings of International Conference on Image and Graphics*. 2007: 683–688.

[41] Mu Y., Yan S., Liu Y., et al. Discriminative local binary patterns for human detection in personal album[C]. *In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, 2008.

[42] Wang X., Han T. X., Yan S.. An HOG–LBP human detector with partial occlusion handling[C]. *In: Proceedings of IEEE International Conference on Computer Vision*, 2009: 32–39.

[43] Ren X., Ramanan D.. Histograms of sparse codes for object detection[C]. *In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, 2013: 3246–3253.

[44] Dollár P., Tu Z., Perona P., et al. Integral channel features[C]. *In: Proceedings of British*

Machine Vision Conference, 2009: 1–11.

[45] Dollár P., Belongie S. J., Perona P.. The fastest pedestrian detector in the west[C]. *In: Proceedings of British Machine Vision Conference*, 2010: 1–11.

[46] Nam W., Dollár P., Han J. H.. Local decorrelation for improved pedestrian detection[C]. *In: Proceedings of Advances in Neural Information Processing Systems*, 2014: 424–432.

[47] Tuzel O., Porikli F., Meer P.. Pedestrian detection via classification on Riemannian manifolds[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2008, 30(10): 1713–1727.

[48] Sermanet P., Kavukcuoglu K., Chintala S., et al. Pedestrian detection with unsupervised multi–stage feature learning[C]. *In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, 2013: 3626–3633.

[49] Lim J. J., Zitnick C. L., Dollár P.. Sketch tokens: A. learned mid–level representation for contour and object detection[C]. *In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, 2013: 3158–3165.

[50] Zhang S., Bauckhage C., Cremers A. B.. Informed haar–like features improve pedestrian detection[C]. *In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, 2014: 947–954.

[51] Gao W., Chen X., Ye Q., et al. Pedestrian detection via part–based topology model [C]. *In: Proceedings of IEEE International Conference on Image Processing*. 2012: 445–448.

[52] Felzenszwalb P. F., Girshick R. B., Mcallester D. A., et al. Object detection with discriminatively trained part–based models[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2010, 32(9): 1627–1645.

[53] Sabzmeydani P., Mori G.. Detecting pedestrians by learning shapelet features [C]. *In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, 2007.

[54] Wu B., Nevatia R.. Detection of multiple, partially occluded humans in a. single image by bayesian combination of edgelet part detectors[C]. *In: Proceedings of IEEE International Conference on Computer Vision*, 2005: 90–97.

[55] Zhao L., Thorpe C. E.. Stereo- and neural network-based pedestrian detection[J]. *IEEE Transaction on Intelligent Transportation Systems*, 2000, 1(3): 148–154.

[56] Keller C. G., Enzweiler M., Rohrbach M., et al. The benefits of dense stereo for pedestrian detection[J]. *IEEE Transaction on Intelligent Transportation Systems*, 2011, 12(4): 1096–1106.

[57] Nishiyama M., Seki A., Watanabe T.. Stereo-based pedestrian detection using two–stage classifiers[C]. *In: Proceedings of International Conference on Machine Vision Applications*. 2011: 520–523.

[58] Dalal N., Triggs B., Schmid C.. Human detection using oriented histograms of flow and

appearance[C]. *In: Proceedings of European Conference on Computer Vision*, 2006: 428–441.

[59] Yan J., Zhang X., Lei Z., et al. Robust multi-resolution pedestrian detection in traffic scenes[C]. *In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, 2013: 3033–3040.

[60] Spinello L., Arras K. O.. People detection in RGB-D data[C]. *In: Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2011: 3838–3843.

[61] Bo L., Lai K., Ren X., et al. Object recognition with hierarchical kernel descriptors [C]. *In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, 2011: 1729–1736.

[62] Walk S., Majer N., Schindler K., et al. New features and insights for pedestrian detection[C]. *In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, 2010: 1030–1037.

[63] Park D., Zitnick C. L., Ramanan D., et al. Exploring weak stabilization for motion feature extraction[C]. *In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, 2013: 2882–2889.

[64] Ding Y., Xiao J.. Contextual boost for pedestrian detection[C]. *In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, 2012: 2895–2902.

[65] Chen G., Ding Y., Xiao J., et al. Detection evolution with multi–order contextual co–occurrence[C]. *In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, 2013: 1798–1805.

[66] Girshick R. B., Iandola F. N., Darrell T., et al. Deformable part models are convolutional neural networks[C]. *In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, 2015: 437–446.

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

[68] Wang M., Wang X.. Automatic adaptation of a. generic pedestrian detector to a. specific traffic scene[C]. *In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, 2011: 3401–3408.

[69] Wang X., Wang M., Li W.. Scene–specific pedestrian detection for static video surveillance[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2014, 36(2): 361–374.

[70] Zeng X., Ouyang W., Wang M., et al. Deep learning of scene–specific classifier for pedestrian detection[C]. *In: Proceedings of European Conference on Computer Vision*, 2014: 472–487.

[71] Vazquez D., Lopez A. M., Marin J., et al. Virtual and real world adaptation for pedestrian

detection[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2014, 36(4): 797–809.

[72] Xu J., Ramos S., Vázquez D., et al. Domain adaptation of deformable part–based models[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2014, 36 (12): 2367–2380.

[73] Kuznetsova A., Ju Hwang S., Rosenhahn B., et al. Expanding object detector's horizon: incremental learning framework for object detection in videos[C]. *In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, 2015: 28–36.

[74] Gaidon A., Zen G., Rodríguez–Serrano J. A.. Self–learning camera: Autonomous adaptation of object detectors to unlabeled video streams[J]. *CoRR*, 2014, abs/1406.4296.

[75] Mao Y., Yin Z.. Training a. scene–specific pedestrian detector using tracklets[C]. *In: Proceedings of IEEE Winter Conference on Applications of Computer Vision*, 2015: 170–176.

[76] Kalal Z., Mikolajczyk K., Matas J.. Tracking–learning–detection[J]. *IEEE transactions on pattern analysis and machine intelligence*, 2012, 34(7): 1409–1422.

[77] Kwak S., Cho M., Laptev I., et al. Unsupervised object discovery and tracking in video collections[C]. *In: Proceedings of IEEE International Conference on Computer Vision*, 2015: 3173–3181.

[78] Wu B., Nevatia R.. Improving part based object detection by unsupervised, online boosting[C]. *In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, 2007: 1–8.

[79] Xiao F., Lee Y. J.. Track and segment: An iterative unsupervised approach for video object proposals[C]. In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, 2016: 933–942.

[80] Dollár P., Zitnick C. L.. Structured forests for fast edge detection[C]. *In: Proceedings of IEEE International Conference on Computer Vision*, 2013: 1841–1848.

[81] Lindeberg T.. Edge detection and ridge detection with automatic scale selection[J]. *International Journal of Computer Vision*, 1998, 30(2): 117–156.

[82] Xie S., Tu Z.. Holistically–nested edge detection[C]. *In: Proceedings of IEEE International Conference on Computer Vision*, 2015: 1395–1403.

[83] Arbelaez P., Maire M., Fowlkes C. C., et al. Contour detection and hierarchical image segmentation[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2011, 33(5): 898–916.

[84] Long J., Shelhamer E., Darrell T.. Fully convolutional networks for semantic segmentation[C]. *In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, 2015: 3431–3440.

[85] Lecun Y., Boser B. E., Denker J. S., et al. Backpropagation applied to handwritten zip code recognition[J]. *Neural Computation*, 1989, 1(4): 541–551.

[86] He K., Zhang X., Ren S., et al. Deep residual learning for image recognition[C]. *In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, 2016: 770–778.

[87] Huang G., Liu Z., Van Der Maaten L., et al. Densely connected convolutional networks[C]. *In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, 2017: 2261–2269.

[88] Lee C., Xie S., Gallagher P. W., et al. Deeply–supervised nets[C]. *In: Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*. 2015.

[89] Liu Y., Cheng M., Hu X., et al. Richer convolutional features for edge detection [C]. *In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, 2017: 5872–5881.

[90] Sebastian T. B., Klein P. N., Kimia B. B.. Recognition of shapes by editing their shock graphs[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2004, 26(5): 550–571.

[91] Trinh N. H., Kimia B. B.. Skeleton Search: Category–specific object recognition and segmentation using a. skeletal shape model[J]. *International Journal of Computer Vision*, 2011, 94(2): 215–240.

[92] Teo C. L., Fermüller C., Aloimonos Y.. Detection and segmentation of 2d curved reflection symmetric structures[C]. *In: Proceedings of IEEE International Conference on Computer Vision*, 2015: 1644–1652.

[93] Fu H., Cao X., Tu Z., et al. Symmetry constraint for foreground extraction[J]. *IEEE Transactions on Cybernetics*, 2014, 44(5): 644–654.

[94] Lee T. S. H., Fidler S., Dickinson S. J.. Learning to combine mid–level cues for object proposal generation[C]. *In: Proceedings of IEEE International Conference on Computer Vision*, 2015: 1680–1688.

[95] Zhang Z., Shen W., Yao C., et al. Symmetry–based text line detection in natural scenes[C]. *In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, 2015: 2558–2567.

[96] Lam L., Lee S., Suen C. Y.. Thinning methodologies –. A. comprehensive survey[J]. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 1992, 14(9): 869–885.

[97] Saha P. K., Borgefors G., Di Baja G. S.. A. survey on skeletonization algorithms and their applications[J]. *Pattern Recognition Letters*, 2016, 76: 3–12.

[98] Liu J., Slota G.. Zheng G.. et al. Symmetry Detection from RealWorld Images Competition 2013: Summary and Results[C]. *In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition Workshops*, 2013: 200–205

[99] Tsogkas S., Kokkinos I.. Learning–based symmetry detection in natural images[C]. *In: Proceedings of European Conference on Computer Vision*, 2012: 41–54.

[100] Shen W., Zhao K., Jiang Y., et al. Object skeleton extraction in natural images by fusing scale–associated deep side outputs[C]. *In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, 2016: 222–230.

[101] Shen W., Bai X., Hu Z., et al. Multiple instance subspace learning via partial random projection tree for local reflection symmetry in natural images[J]. *Pattern Recognition*, 2016, 52: 306–316.

[102] Simonyan K., Zisserman A.. Very deep convolutional networks for large–scale image recognition[C]. International Conference on Learning Representations. 2015.

[103] Freund Y., Schapire R. E.. Experiments with a. new boosting algorithm[C]. *In: Proceedings of International Conference on Machine Learning*, 1996: 148–156.

[104] De Santis D.. High–order linear and non–linear residual distribution schemes for turbulent compressible flows[J]. *Computer Methods in Applied Mechanics and Engineering*, 2015, 285: 1–31.

[105] Dollár P., Zitnick C. L.. Fast edge detection using structured forests[J]. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 2015, 37(8): 1558–1570.

[106] Levinshtein A., Sminchisescu C., Dickinson S. J.. Multiscale symmetric part detection and grouping[J]. *International Journal of Computer Vision*, 2013, 104(2): 117–134.

[107] Lee T. S. H., Fidler S., Dickinson S. J.. Detecting curved symmetric parts using a. deformable disc model[C]. *In: Proceedings of IEEE International Conference on Computer Vision*. 2013: 1753–1760.

[108] [108] Widynski N., Moevus A., Mignotte M.. Local symmetry detection in natural images using a. particle filtering approach[J]. *IEEE Transaction on Image Processing*, 2014, 23(12): 5309–5322.

[109] Ren S., He K., Girshick R. B., et al. Faster R–CNN: towards real–time object detection with region proposal networks[C]. *In: Proceedings of Advances in Neural Information Processing Systems*, 2015: 91–99.

[110] Shen W., Bai X., Hu R., et al. Skeleton growing and pruning with bending potential ratio[J]. *Pattern Recognition*, 2011, 44(2): 196–209.

[111] Canny J. F.. A. computational approach to edge detection[J]. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 1986, 8(6): 679–698.

[112] Shen W., Wang X., Wang Y., et al. Deepcontour: A. deep convolutional feature learned by positive–sharing loss for contour detection[C]. *In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, 2015: 3982–3991.

[113] Felzenszwalb P. F., Huttenlocher D. P.. Efficient graph–based image segmentation[J]. *International Journal of Computer Vision*, 2004, 59(2): 167–181.

[114] Xiao Y., Lu C., Tsougenis E., et al. Complexity–adaptive distance metric for object proposals generation[C]. *In: Proceedings of IEEE International Conference on Computer Vision and*

Pattern Recognition, 2015: 778–786.

[115] Chen X., Ma H., Wang X., et al. Improving object proposals with multithresholding straddling expansion[C]. *In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, 2015: 2587–2595.

[116] Everingham M., Eslami S. M. A., Gool L. J. V., et al. The pascal visual object classes challenge: A. retrospective[J]. *International Journal of Computer Vision*, 2015, 111(1): 98–136.

[117] Alexe B., Deselaers T., Ferrari V.. Measuring the objectness of image windows[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2012, 34(11): 2189–2202.

[118] Arbeláez P. A., Pont–Tuset J., Barron J. T., et al. Multiscale combinatorial grouping[C]. *In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, 2014: 328–335.

[119] Carreira J., Sminchisescu C.. CPMC: automatic object segmentation using constrained parametric min–cuts[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2012, 34(7): 1312–1328.

[120] Endres I., Hoiem D.. Category–independent object proposals with diverse ranking[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2014, 36(2): 222–234.

[121] Humayun A., Li F., Rehg J. M.. RIGOR: reusing inference in graph cuts for generating object regions[C]. *In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, 2014: 336–343.

[122] Manen S., Guillaumin M., Gool L. J. V.. Prime object proposals with randomized prim's algorithm[C]. *In: Proceedings of IEEE International Conference on Computer Vision*. 2013: 2536–2543.

[123] Rahtu E., Kannala J., Blaschko M. B.. Learning a. category independent object detection cascade[C]. *In: Proceedings of IEEE International Conference on Computer Vision*. 2011: 1052–1059.

[124] Rantalankila P., Kannala J., Rahtu E.. Generating object segmentation proposals using global and local search[C]. *In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, 2014: 2417–2424.

[125] Chan T., Jia K., Gao S., et al. PCANET: A. simple deep learning baseline for image classification?[J]. *IEEE Transaction on Image Processing*, 2015, 24(12): 5017–5032.

[126] Dollár P., Wojek C., Schiele B., et al. Pedestrian detection: A. benchmark [C]. *In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, 2009: 304–311.

[127] Benenson R., Mathias M., Timofte R., et al. Pedestrian detection at 100 frames per second[C]. *In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, 2012: 2903–2910.

[128] Benenson R., Mathias M., Tuytelaars T., et al. Seeking the strongest rigid detector[C]. *In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, 2013: 3666–3673.

[129] http://www.vision.caltech.edu/Image–Datasets/CaltechPedestrians/.

[130] Hattori H., Boddeti V. N., Kitani K. M., et al. Learning scene–specific pedestrian detectors without real data[C]. *In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, 2015: 3819–3827.

[131] Misra I., Shrivastava A., Hebert M.. Watch and learn: Semi–supervised learning of object detectors from videos[C]. *In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, 2015: 3593–3602.

[132] Yu C. J., Joachims T.. Learning structural svms with latent variables[C]. *In: Proceedings of International Conference on Machine Learning*. 2009: 1169–1176.

[133] Zhu X., Goldberg A. B.. Introduction to semi–supervised learning[M]. 2009.

[134] Shi J., Tomasi C.. Good features to track[C]. *In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, 1994: 593–600.

[135] Chang C., Lin C.. LIBSVM: A. library for support vector machines[J]. *ACM Transaction on Intelligent System and Technology*, 2011, 2(3): 27:1–27:27.

[136] Knuth D.. Sorting and searching in the art of programming[J]. Addison Wesley, 1973, 3: 506.

[137] Ferryman J., Shahrokni A.. Pets2009: Dataset and challenge[C]. *IEEE International Workshop on Performance Evaluation of Tracking and Surveillance*. 2009: 1–6.

[138] Benfold B., Reid I. D.. Stable multi–target tracking in real–time surveillance video[C]. *In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, 2011: 3457–3464.

致　谢

时间过得很慢，经历了遍寻研究方向的迷茫、实验结果不及人意的沮丧、两度延期的苦楚。时间过得也很快，学期末总结中带走的一个个半年，赶论文截止日期时迅速溜走的日日夜夜。在博士这七年里，我扎根于中国科学院大学模式识别与智能系统开发实验室，先后访问清华宽带网数字媒体实验室、芬兰奥卢大学计算机视觉研究中心。七年的坚持，终是守的云开见日出。得到的不仅是一篇博士学位论文，而且是综述、发现、思考和解决问题的能力和对计算机视觉更深刻的认识；不仅有自己的历练，而且有老师的指导、同学朋友的帮助和家人的关怀支持。

本论文的研究工作是在叶齐祥教授、焦建彬教授、韩振军副教授和秦飞副教授的悉心指导下完成的。叶老师亦师亦友，科研上给予我深刻的见解，论文写作上给与我悉心的指导，生活里给与我积极的鼓励。叶老师对科研的执着和热爱值得我一生学习。焦老师为人谦和、治学严谨，他给了实验室每个人科研的自由度并提供指导，在生活上对我们也关怀备至。韩老师和秦老师与我年龄相仿，他们思维缜密、工作认真、经历丰富、平易近人，给我诸多帮助与启迪。

真诚地感谢叶齐祥教授、焦建彬教授、韩振军副教授和秦飞副教授，也真诚地感谢那些在我学习和生活中提供过无私帮助与支持的其他老师、同学以及亲人。

感谢芬兰奥卢大学计算机视觉研究中心的陈杰老师和赵国英教授。在奥卢大学访问的一年里，陈老师耐心的讨论，坚实的鼓励，才让我顶住压力在即将延期的那一年投中顶会论文，增强了我的科研自信。赵老师和蔼可亲，无论科研上还是生活上，都给我很多建议和帮助。也要感谢在奥卢遇到其他朋友，在国外的这一年，他们就如同亲人一般，让我在接近北极的芬兰也感到温暖。

感谢在清华宽带网数字媒体实验室的季向阳教授和汪启扉博士。在刚准备踏入科研时遇到了季老师和汪博士，他们的科研成果让我敬仰，是他们让我知道了顶会、汇刊，也是他们的悉心指导让我完成了第一轮科研训练。

感谢北航的张宝昌老师和李琳同学，是他们帮我从电气工程及其自动化走上了计算机视觉的道路。

感谢我的小伙伴们。我觉得自己真的很幸运。有那么几个初中高中的铁哥们儿。有那么一队基本都还在北京的大学同学。有那么一群因为实验室搬到怀柔雁栖湖而整天一起泡在实验室的师兄师姐师弟师妹。每个团结亲和的集体都给我无尽的享受和温暖。

感谢父母、叔叔姑姑，每每听到远在陕西的他们对我说要吃好学好的时候都很感动。这朴实的话是他们无私的关爱、鼓励和支持。感谢妹妹和妹夫，正是她们让我能“父母在能远游”。特别感谢我的妻子陈茜对我的理解和支持，八年相候，不离不弃，同甘共苦，在即将毕业之际终能执子之手。

感谢参加开题、中期和毕业答辩的各位指导老师专家，你们丰富的经验和细致的指导，对论文方向和研究进度的指点给我的研究工作带来了巨大的帮助。

借着写学位论文的机会回顾过去，才发现我遇到了如此多的美好的人。科研不易，有了你们才让我的科研多姿多彩。最后，再次向在学习、工作和生活中给予过自己关心、支持与鼓励的所有老师、同学、朋友们表示最诚挚的谢意！

柯　炜

2018 年 2 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：

2007 年 09 月——2011 年 07 月，在北京航空航天大学自动化科学与电气工程学院获得学士学位

2015 年 10 月——2016 年 10 月，在芬兰奥卢大学访问

2011 年 09 月——2018 年 03 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

获奖情况：2017 年获得中国科学院院长优秀奖

2016 年获得研究生国家奖学金

2012 年和 2015 年获得中国科学院大学三好学生

已发表（或正式接受）的学术论文：

一作论文：

[1] **Wei Ke**, Jie Chen, Jianbin Jiao, Guoying Zhao, and Qixiang Ye. SRN: Side-output Residual Network for Object Symmetry Detection in the Wild[C]. *In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition (CVPR, Oral)*, 2017: 1068-1076. (EI)

[2] **Wei Ke,** Jie Chen, and Qixiang Ye. Deep Contour and Symmetry Scored Object Proposal[J]. *Pattern Recognition Letters*, Elsevier, 2018. (SCI, 已接收)

[3] **Wei Ke**, Tianliang Zhang, Jie Chen, Fang Wan, Qixiang Ye and Zhenjun Han. Texture Complexity based Redundant Regions Ranking for Object Proposal[C]. *In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition Workshop (CVPRW)*, 2016:354-365. (EI)

[4] **Wei Ke**, Yao Zhang, Pengxu Wei, Qixiang Ye, and Jianbin Jiao. Pedestrian Detection via PCA Filters Based Convolutional Channel Features [C]. *In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015:1394-1398. (EI)

合作论文：

[1] Qixiang Ye, Tianliang Zhang, **Wei Ke**, et al., Self-learning Scene-specific Pedestrian Detectors

using a Progressive Latent Model [C]. *In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017: 509-518. (EI)

[2] Chang Liu, **Wei Ke**, Jianbin Jiao, and Qixiang Ye. RSRN: Rich Side-output Residual Network for Medial Axis Detection [C]. *In: Proceedings of IEEE International Conference on Computer Vision Workshop (ICCVW)*, 2017:1739-1743.

[3] Qing Liu, Beiji Zou, Jie Chen, **Wei Ke**, et al.. A Location-to-Segmentation Strategy for Automatic Exudate Segmentation in Colour Retinal Fundus Images. *Computerized Medical Imaging and Graphics*, vol 55, 2017. 78-86,

[4] Xiaogang Chen, Pengxu Wei, **Wei Ke**, Qixiang Ye, and Jianbin Jiao. Pedestrian Detection with Deep Convolutional Neural Network [C]. *In: Proceedings of Asian Conference on Computer Vision (ACCV) Workshop*, 2014:354-365. (EI)

[5] Liguo Zhang, **Wei Ke**, Qixiang Ye, and Jianbin Jiao. A Novel Laser Vision Sensor for Weld Line Detection on Wall-climbing Robot. *Optics and Laser Technology*, vol 66, 2014. 69-79.

[6] Wei Yang, Liguo Zhang, **Wei Ke**, Ce Li, and Jianbin Jiao. Minimum entropy models for laser line extraction. *In: Proceedings of International Conference on Computer Analysis of Images and Patterns (CAIP)*, 2013. 499-506.

[7] Liguo Zhang, **Wei Ke**, Zhenjun Han, and Jianbin Jiao. A Cross Structured Light Sensor for Weld Line Detection on Wall-climbing Robot. *In: Proceedings of IEEE International Conference on Mechatronics and Automation (ICMA)*, 2013. 1179-1184.

在审论文：

[1] **Wei Ke**, Jie Chen, Jianbin Jiao, Guoying Zhao, and Qixiang Ye. SRN: Side-output Residual Network for Object Symmetry Detection and Beyond [J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2017. (SCI)