

中国科学院自动化研究所

研究生学位论文中期考核报告

论文题目：物体检测中的锚点框研究

专 业：模式识别与智能系统

研究方向：物体检测

姓 名：张士峰

学 号：201518014628079

培养层次：☒博士 ☐硕士

博士攻读方式：☐硕博连读 ☒直接攻博 ☐普通招考

导师姓名：李子青

所属部门：模式识别国家重点实验室

考核日期：2019 年 11 月 21 日

目 录

一、研究背景与意义	1
二、国内外研究概述	3
1. 传统方法	3
2. 深度学习方法	4
三、学位论文撰写提纲	5
1. 绪论	5
2. 物体检测的研究现状	6
3. 物体检测中锚点框的设计匹配	6
4. 物体检测中锚点框的关系学习	6
5. 物体检测中锚点框的高效预测	7
6. 总结与展望	8
四、学位论文主要创新成果	8
1. 锚点框的设计匹配	8
A) 高精度人脸检测算法 SFD	8
B) CPU 实时人脸检测算法 FACEBOXES	10
2. 锚点框的关系学习	12
A) 分块聚合行人检测算法 OR-CNN	12
B) 人头人体联合检测算法 JOINTDET	13
3. 锚点框的高效预测	14
A) 取长补短的物体检测算法 REFINEDET	14
B) 自适应训练样本选取算法 ATSS	16
4. 学位论文工作进度安排	18
5. 已取得的阶段性成果	18
6. 课程主要完成情况	21
7. 其他	23
主要参考文献	23

物体检测中的锚点框研究

一、 研究背景与意义

随着大数据时代的到来，时时刻刻都有大量的视频和图像数据产生，如此庞大的数据量远远超越了人类的分析能力，导致人工地去发现和查找特定目标（如人脸、行人、车辆等）变得不可能，使得人们淹没在了信息的海洋中。如何利用计算机来智能且高效地处理这些信息，以更好的服务社会成为了目前炙手可热的研究方向，其中，基于图像和视频的物体检测技术是不可或缺的重要一环。

物体检测是一种使计算机能够在给定图像或视频中自动找到给定类别的物体，并判断物体的类别、位置、大小及置信度的技术。简单来说，即物体的定位（where）与识别（what）。其中，物体定位负责在给定图像或视频中找到物体所在的位置，输出包围物体的矩形框，而物体识别负责判断该位置物体的类别，输出一系列置信度来表明各类的可能性。确定了物体的位置和类别后，接下来就可以应用到各种各样的任务中。因此，从研究角度看，物体检测是计算机视觉的根本问题之一，是很多高层计算机视觉任务的基础，比如目标跟踪、图像描述、人脸识别、事件检测、情感理解、行人再辨识、场景理解等，而这些任务正是下一代人机交互和高级人工智能的基础。

从应用角度看，物体检测已经表现出广泛的应用需求和前景。外貌美化、人脸装饰、笑脸快门、辅助驾驶、图像搜索等技术已经在不经意间默默提升了我们的生活质量。在不远的将来，更加善解人意的人工智能类应用更是值得我们期待。想象一下，车辆可以在撞到行人之前自动停下来，相机可以自动根据环境来调整参数设置，智能机器人可以自动地与真实的世界进行交互。所有这些应用与改变中，都会有物体检测作为核心技术在背后的贡献，下面列举一些物体检测技术应用潜力巨大的领域。

- 辅助驾驶。众多汽车已经应用了辅助驾驶技术，旨在给予驾驶者以危险提醒，对于行人、车辆的检测可以有效避免恶性交通事故的发生；对于路牌，交通标示牌的检测则可以帮助车辆更加安全稳定地行驶。这个应用的核心问题是物体检测技术的鲁棒性，如何在确保召回率的情况下提高正确率，使行人、车辆、路牌检测能真正让人放心使用。
- 自动驾驶。物体检测是自动驾驶系统的核心组成部分，它能自动检测车前方的人、车、路标、车道线、红绿灯以及其他的障碍物等，并且随后做出相应的反馈，实现自动安全驾驶。各大汽车厂商，包括谷歌、丰田和特斯拉等厂商都在研究自动驾驶系统，其中，谷歌的无人车已经在加州安全行驶了上百万公里。目前的无人车还非常依赖成本非常高的雷达系统，而鲁棒可靠的物体检测系统有可能帮助降低这些额外的硬件的成本。
- 智能视频监控。随着监控摄像头的越来越普及，如何有效地分析这些视频成了一个需要解决的关键问题。传统的人工分析方法越来越不能适应日益增多的数据，在这种情况下，智能视频监控应用而生。其中，物体检测技术是智能视频监控的基础，它可以提供人脸、行人和车辆等的位置和属性类别标签，从而为后续的生物特征识别、属性识别以及行为识别等提供基础。智能视频监控场景多种多样，给物体检测技术带来了巨大的挑战。目前，有很多创业公司都在做智能视频监控系统，比例商汤、旷视、依图等。
- 智能家居。随着生活节奏的日益加快，人们很少有时间管理日常生活的方方面面。在不久的将来，如果家里能有一位智能管家，知道电冰箱水果和食物的储备，并在过期之前提醒尽快食用，吃完之前自动订购；知道电费和水电费剩余情况，费用即将使用完之前自动缴费，就不会出现突然断电、断水。而这一切的核心技术就是物体检测，不需要人工干预，所有的一切都由机器自动完成，极大地方便了日常生活。
- 人机交互。随着科技的发展，人类已经逐步脱离了键盘和鼠标为主的人机接口方式，向更自然、更拟人化的人机交互迈进，包括手势、肢体动作、语言

等自然的交互方式。在这些智能化、人性化的人机交互方式中，物体检测扮演着关键角色，因为要让计算机理解人的想法，需要对人体进行运动分析，其中最基础也是最关键的一步就是对其人脸和手指的检测与跟踪。

- 智能机器人：对于一些危险的行业，由于人类易疲劳、易疏忽，难免会发生一些意外事件，造成人员的重大损伤，如果能用智能机器人取代，可以大幅度降低意外发生的概率，同时也能最大程度地降低损伤。此外有些人类很难到达的地方，可以利用智能移动机器人来代替人类完成相应的操作，这些机器人也能自动检测识别前方的物体，随后做出相应的行为等。

综上所述，物体检测有着重要的研究地位以及广泛的应用价值，对物体检测的研究具有非常重要的意义，物体检测技术的发展与创新值得期待。

二、国内外研究概述

物体检测是计算机视觉领域的基本任务之一，学术界已有将近二十年的研究历史。随着深度学习技术的火热发展，物体检测算法也从基于手工特征的传统算法转向了基于深度神经网络的检测方法。

1. 传统方法

传统方法的主要流程是，在图像上进行密集地窗口滑动，对每个滑动窗口提取手工特征，送给人为设计的分类器进行分类。传统方法使用大小固定的滑动窗口，所以需要图像金字塔来解决物体的尺度问题。传统方法的核心在于手工设计特征和分类器。在手工特征方面，从早期的 Viola 和 Jones 的 Haar-Like 特征[1]，Dalal 的 HOG 特征[2]，Dollár 的积分通道特征（ICF）[3]到最近的集合通道特征（ACF）[4]，Shanshan Zhang 的 Informed Haar[5]特征和滤波通道特征（Checkboard）[6]，从计算速度和检测的准确率上都有了较大的提高。在手工分类器方面，大多采用级联的 Adaboost[7-9]、SVM[10]、Decision Tree[11]、Random Forest[11]等机器学习算法。传统物体检测方法的主要研究点集中在对

特征的改进、对特征金字塔计算的加速、对分类器的改进。但传统方法存在着一些缺点：手工设计的特征通用性不强、固定滑窗方式不能很好地适应物体多样性、密集滑窗存在着大量的冗余计算、各步骤独立容易得到局部最优。虽然诸多缺点导致了传统方法的性能受限，但是它们的思想对后续的深度学习有着重要的指导意义。

2. 深度学习方法

在 2012 年的 ImageNet 比赛中，Geoffrey Hinton 的两个学生使用基于深度学习的方法[16]获得了冠军，其精度比传统方法高不少，自此深度学习时代到来。起初深度学习只是在分类上有非常明显的提升，后来由于深度学习较强的特征表达能力，也带动了物体检测的发展，使得检测性能大幅度被提高。目前，基于深度学习的物体检测方法大致分为两类：二步检测法和一步检测法。二步检测法由两个步骤组成，第一步产生少量候选区域位置，第二步使用卷积网络对每个候选区域进行分类和回归，得到最终的检测结果。一步检测法则直接使用卷积网络对初始锚点框进行分类和回归，得到检测结果。

二步检测法的基本思想是先得到候选区域，再对候选区域进行分类和回归。此方法检测精度较高，但检测速度偏慢，代表作有 Faster R-CNN[22]、R-FCN[23]、FPN[24]。Faster R-CNN 是二步检测方法的奠基性工作，使得检测任务能够在在一个网络中端到端地完成。在此之前的 R-CNN[17]、SPPNet[20]和 Fast R-CNN[21]，它们的候选区域都是由 Selective Search[18]、Edgebox[19]等底层算法单独地离线计算出来。Faster R-CNN 的主要贡献是提出了一个区域生成网络（Region Proposal Network, RPN）来专门产生候选区域，与后续的检测网络 Fast R-CNN 一起端到端地训练。R-FCN 使第二级尽量地共享后续子网络来提高检测的速度，通过 PSROI Pooling 层来保持检测对于物体的平移变化性，同时也可以去掉之前的逐区域操作子网络，使得检测的速度大大提升。FPN 在增加网络深度、获取更丰富语义信息的同时从浅层特征图中获取更丰富且高分辨率的图像特征，这使得这种网络结构在实际应用中表现出优异的性能。物体检测领域进入深度学习时代

后，二步检测法就开始在本领域中占主导地位，从最初的 R-CNN、SPPNet、Fast R-CNN、Faster R-CNN，到后来的 R-FCN、FPN，物体检测精度越来越高，一直在常用数据集（PASCAL VOC[25]、MS COCO[26]）的榜单中处于领先地位。在奠基性工作 Faster R-CNN 之后，该类方法的检测流程基本没有本质性变化，研究点主要集中在框架改进[27]、训练策略[28]、卷积形式[29]、特征融合[30]等。

一步检测法没有候选区域生成阶段，直接根据图像获得检测结果。此方法检测速度较快，检测精度不够高，代表作有 YOLO[31]、SSD[32]、RetinaNet[33]。YOLO[31]是一步法的开山之作，它将物体检测任务表述成一个统一的、端到端的回归问题，并且以只处理一次图像同时得到位置和分类而得名。SSD[32]是一步法的集大成者，检测精度与二步法接近的同时，拥有比二步法快一个数量级的速度。后续的一步法工作大多是基于 SSD 改进展开。SSD 相比 YOLO 的不同点在于它同时在多个特征图上做预测，每个特征图上关联了不同尺度和比例的锚点框。SSD 的核心思路是采用不同尺寸不同深度的特征层分别进行检测。RetinaNet[33]提出 Focal Loss 来解决样本类别不平衡的问题，在计算交叉熵损失时引入调制因子，降低那些分类任务完成较好的样本的权重，从而重点关注误分的样本，以此提高占比较低的误分样本在训练时的作用。虽然一步检测算法起步比二步检测算法晚，但是由于其高效性得到了学术界的关注，从最初的 YOLO、SSD，到后来的 YOLOv2、YOLOv3、RetinaNet，在保持高效性的前提下，物体检测精度越来越高，在常用的数据集也取得了不错的成绩。近几年一步检测算法的研究点主要集中在特征增强[34]、架构改进[35]等。

三、学位论文撰写提纲

1. 绪论

介绍物体检测的研究背景和意义，引出锚点框在物体检测中的重要作用以及其研究内容。提出该研究专注于物体检测中锚点框的设计匹配、关系学习、高效预测。

2. 物体检测的研究现状

首先介绍经典的传统物体检测方法,以及它们对后续算法的长远影响和指导意义。随后重点介绍基于深度学习的物体检测方法,它分为基于锚点框的物体检测方法和无需锚点框的物体检测方法。前者包括一步检测法和二步检测法,后者包括基于关键点的方法和基于中心区域的方法。本章节中将对这些基于深度学习的物体检测方法依次地进行归类介绍,总结它们的优缺点。

3. 物体检测中锚点框的设计匹配

基于锚点框的物体检测方法,事先需要预设一系列的锚点框,然后把这些锚点框分配给各个物体,最后对这些锚点进行预测得到最终的检测结果。在这些步骤中,锚点框的设计至关重要,所设计的锚点框是检测的起点,它的好坏决定着一个检测算法的性能上限。当锚点框设计好之后,还需要把这些锚点框分配给不同大小的物体。由于锚点框只设计了几个离散的尺度,而各种各样需要检测的物体有着连续的尺度,这就导致锚点框匹配过程中经常出现不平衡的现象。由于锚点框设计与匹配上存在的一些问题,导致小尺度物体的检测性能不够好。

针对锚点框设计中存在的问题,我们提出了基于有效感受野理论和等比间隔原则的锚点框设计方案,以利用有效感受野、特征金字塔和锚点框密集化操作,合理地设计锚点框。针对锚点框匹配中存在的问题,我们提出了尺度补偿的锚点框匹配策略和分而治之的锚点框匹配策略,以使得各个尺度物体所匹配上的锚点框数目比较均匀,平衡地分配锚点框。本章节所提出的锚点框设计与匹配的方法,能够极大地提高小尺度物体的检测性能。

4. 物体检测中锚点框的关系学习

铺设一系列锚点框对物体进行检测时,并没有把锚点框之间的关系、物体本身的结构关系、物体之间的上下文关系利用起来,这就导致基于锚点框的物体检测算法在检测遮挡的物体时,非常容易出现漏检或虚检。物体检测中存在两类遮挡,一类是类间互遮挡,即某一个类别的物体被其他类别的物体遮住,另一类是类内自遮挡,即某一个类别的物体之间相互遮挡住了。类间互遮挡带来的问题是

会引入噪音信息从而会影响分类,类内自遮挡带来的问题是造成信息丢失混淆并抑制重叠物体。这些问题使得基于锚点框的方法在检测遮挡的物体变得极其困难。

为了解决遮挡物体带来的上述问题,首先我们利用物体本身的结构关系,提出了一个分块遮挡感知的特征融合操作,以消除噪音特征来更好的分类。其次我们利用锚点框之间的关系,提出了一个聚合损失函数,让检测结果变得更加紧凑。最后,我们利用物体之间的上下文关系,提出了联合检测方法,召回被抑制的漏检,同时抑制错误的虚检。本章节所提出的关于锚点框关系学习的方法,能够有效地提升遮挡物体的检测性能。

5. 物体检测中锚点框的高效预测

利用锚点框进行物体检测的思路是一个逐步校正的过程。预设的一些列锚点框是检测结果的起点,通过网络对锚点框的类别和位置进行校正,从而得到最终的检测结果。如果对锚点框只进行一次校正,那这类基于锚点框的算法就是一步检测法。如果对锚点框进行两次或两次以上的校正,那这类基于锚点框的算法就是二步检测法。由于二步检测法利用复杂的逐区域操作网络多校正了几次锚点框,而因此它有着更好的检测精度,而一步检测法有这更快的检测速度。为了结合这两类算法让它们优劣互补,我们巧妙地利用特征金字塔结果,它的上半部分用来做第一次校正,下半部分用来做第二次校正,两部分之间是级联的关系,从而利用全卷积网络让一步检测法有了二步检测法的流程,从而提出了具备一步检测法的速度和二步检测法的精度的全新检测算法。

此外,近期无需锚点框的检测算法又流行了起来,其中基于中心区域的无锚点框算法就是铺了一系列的点来做检测,这跟铺一系列框的算法非常类似。经过仔细探索,我们发现无锚点框的算法和基于锚点框的算法,两者的本质区别就是训练中正负样本的定义不一样,而检测的起点是一个框还是一个点没有影响。鉴于此,我们深入探索了检测中最基本的问题,即如何在训练过程中定义正负样本,提出了一种自适应的正负样本划分方法,该方法不需要任何超参数,并在不增加任何开销的情况下,极大地提高了检测性能。

6. 总结与展望

物体检测领域已经被基于锚点框的方法统治了好几年，我们深入研究了锚点框的机理，并在锚点框的设计匹配、关系学习、高效预测三个方面提出了一系列改进，有效提升了该系列方法在小尺度物体、遮挡物体、精度速度平衡三个方面的检测性能。在未来的工作中，我们希望进一步减少整个网络中的手工参数，以增强基于锚点框检测算法的泛化性。

四、学位论文主要创新成果

1. 锚点框的设计匹配

a) 高精度人脸检测算法 SFD

现有的基于锚点框的物体算法在锚点框的设计和匹配上都存在一些问题，导致小尺度物体的检测性能较差，而小尺度物体普遍存在于人脸检测任务中。为此，我们基于著名一步检测方法 SSD，在锚点框的设计和匹配上进行了相应的改进，提出了 SFD 人脸检测算法（如图 1 所示），大幅度提升了小尺度物体的检测性能。

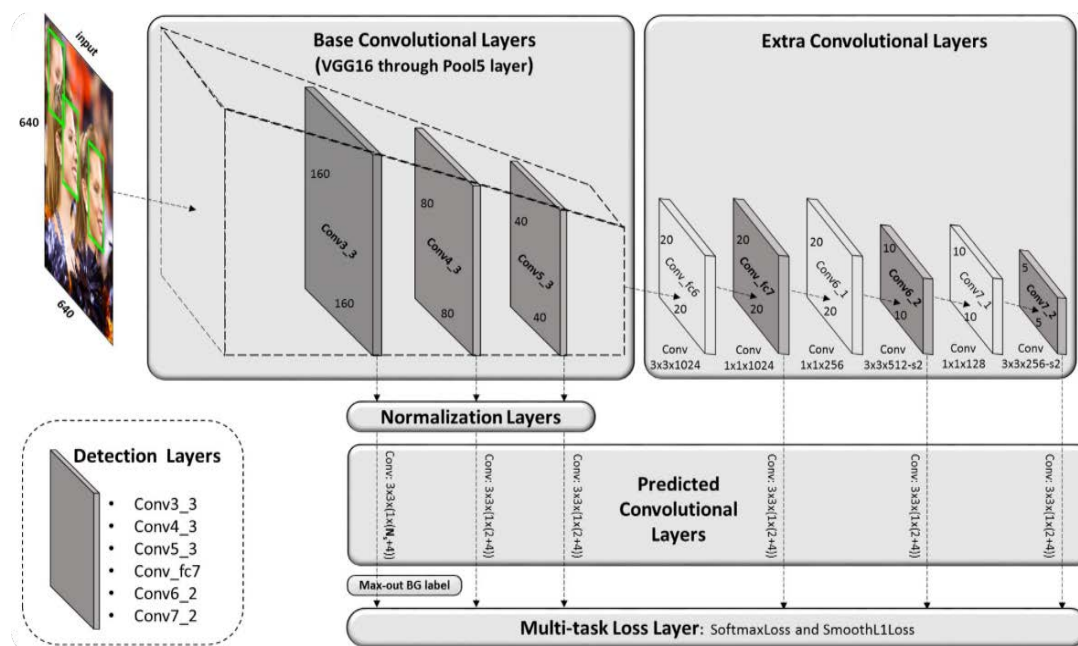


图 1. SFD 的检测框架

在锚点框的设计方面，如图 2(a)所示，我们根据有效感受野理论对锚点框的大小进行设计。在卷积神经网络中，一个神经元有两种感受野。一种是可计算得到的理论感受野，该感受野是理论上影响神经元输出的像素区域，另一种是有效感受野，是真正决定该神经元输出的像素区域。通常来说，有效感受野仅仅是实际感受野很小的一部分。根据以上有效感受野理论，对锚点框的大小进行设计，使得锚点框大小与有效感受野相匹配，可以让网络利用有效感受野中的信息，对锚点框进行分类和回归操作。其次，如图 2(b)所示，锚点框大小的设计还遵循等比间隔原则，它保证了不同大小的锚点框在图像上有着相同的铺设密度，从而能够让不同大小的人脸匹配到数量相等的锚点框，让网络模型公平地对待不同尺度的人脸。

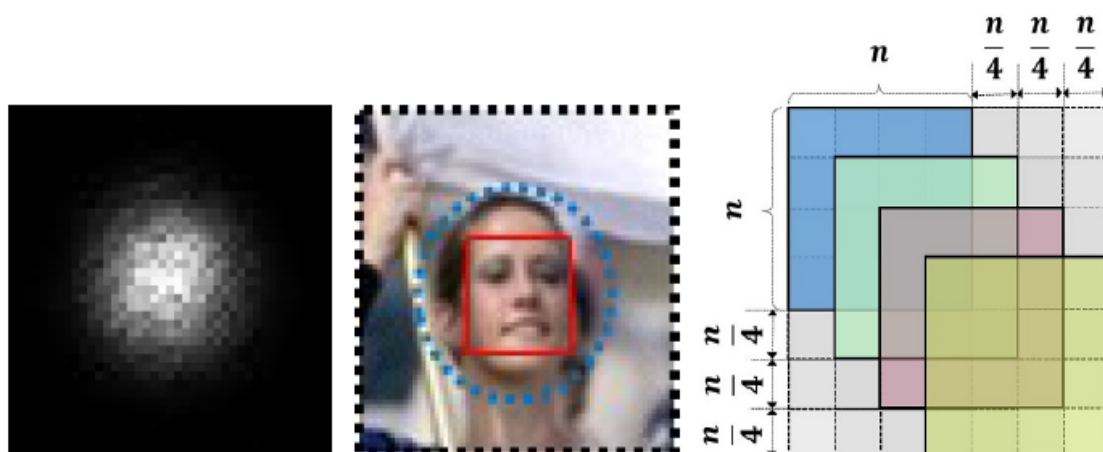


图 2. (a) 有效感受野理论

(b) 等比间隔原则

在锚点框的匹配方面，我们提出尺度补偿的锚点框匹配策略，解决了部分人脸不能匹配到充足的锚点框这一问题。尺度补偿匹配策略由两个步骤组成。第一步，根据标准的锚点框匹配策略，用一个更合理的阈值，对人脸标注框和锚点框进行配对。第二步，进行查缺补漏的尺度补偿操作，对于没有匹配到足够多锚点框的人脸，降低匹配条件让它们匹配上充足的锚点框，从而让各个尺度的人脸能够匹配上数量差不多的锚点框，如图 3 所示，保证了各个尺度之间的公平性。

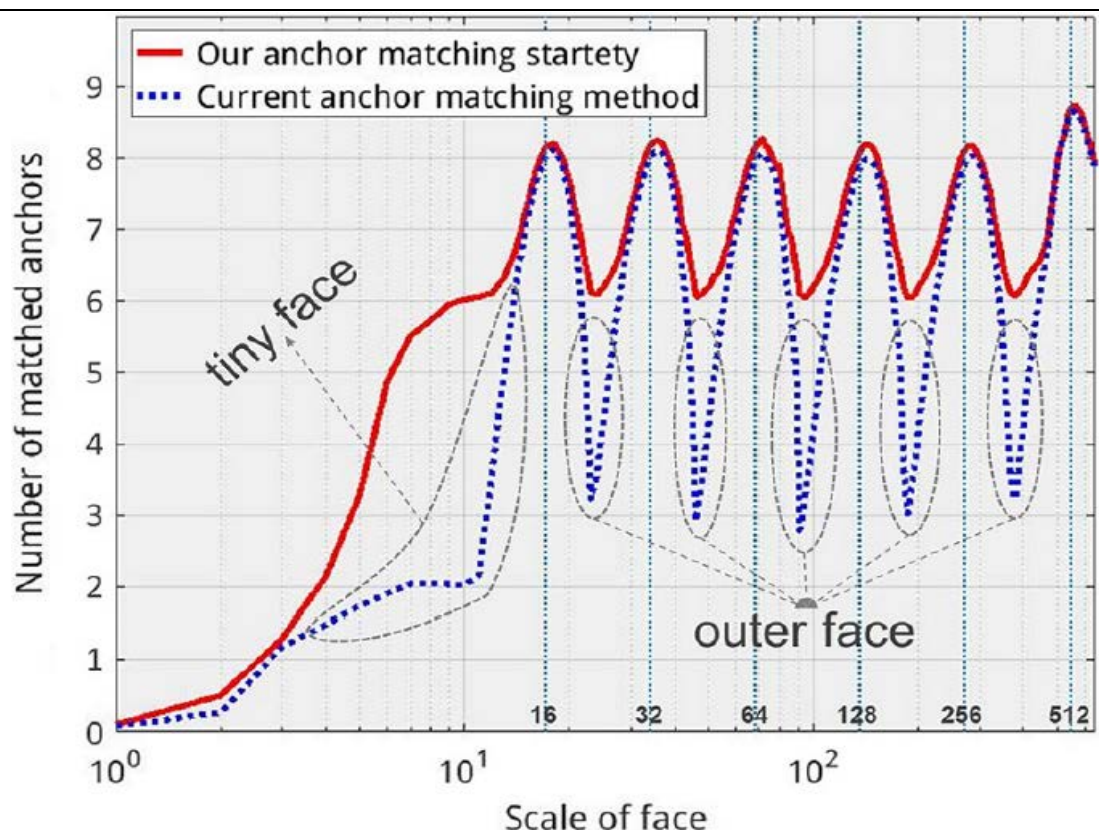


图 3. 尺度补偿的锚点框匹配策略

此外，小尺度锚点框会产生大量的负样本，导致很多的虚检，因此我们针对小尺度锚点框提出了背景标签输出最大化的操作。该研究有效地解决了小尺度人脸难以检测的问题，显著提高了小尺度人脸上的检测性能，在所有主流人脸检测库上，都取得了最好的检测结果，并且能够在 GPU 上以实时的速度进行检测。该研究的相关研究成果已发表在 ICCV 2017 和 IJCV。

b) CPU 实时人脸检测算法 FaceBoxes

为了能够在 CPU 上实时的找出人脸，我们提出了 FaceBoxes 人脸检测算法。该算法设计了一个如图 4 所示高效但轻便的网络架构，这个架构由快速消化网络和多尺度网络组成。快速消化网络利用大小合适的卷积核，快速地降低输入的空间尺寸，并输出较少的通道，以达到 CPU 下实时的检测速度。多尺度网络用来丰富网络的感受野，并把锚点框合理地铺设于适当的卷积层，以保持较高的精度。

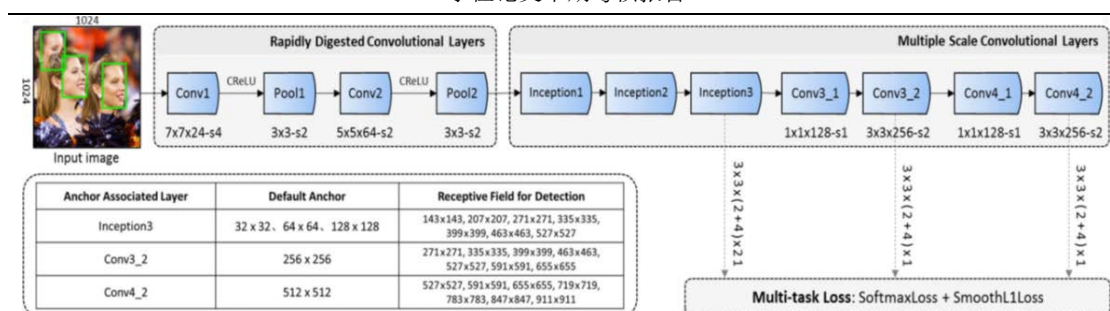


图 4. FaceBoxes 的检测框架

在锚点框的设计方面，由于不同尺度的锚点框，有着不同的铺设密度，铺设较稀疏的锚点框，会导致召回率下降，因此我们提出图 5 所示的锚点框密集化操作，对铺设稀疏的锚点框进行密集化，有效地提高了小尺度人脸的召回率。在锚点框的匹配方面，我们针对不同尺度的人脸设计了不同的匹配策略。对于小尺度人脸，当特定尺度锚点框的中心点在人脸标注框中某一个预设的区域时，就把这个锚点框分配给此人脸。对于大尺度人脸，仍然使用传统基于交叠比的匹配策略。这种分而治之的匹配策略，能够让不同尺度的人脸匹配到充足的锚点框，保证了它们之前的公平性。该研究相关研究成果已发表在 CCBR、IJCB、Neurocomputing。

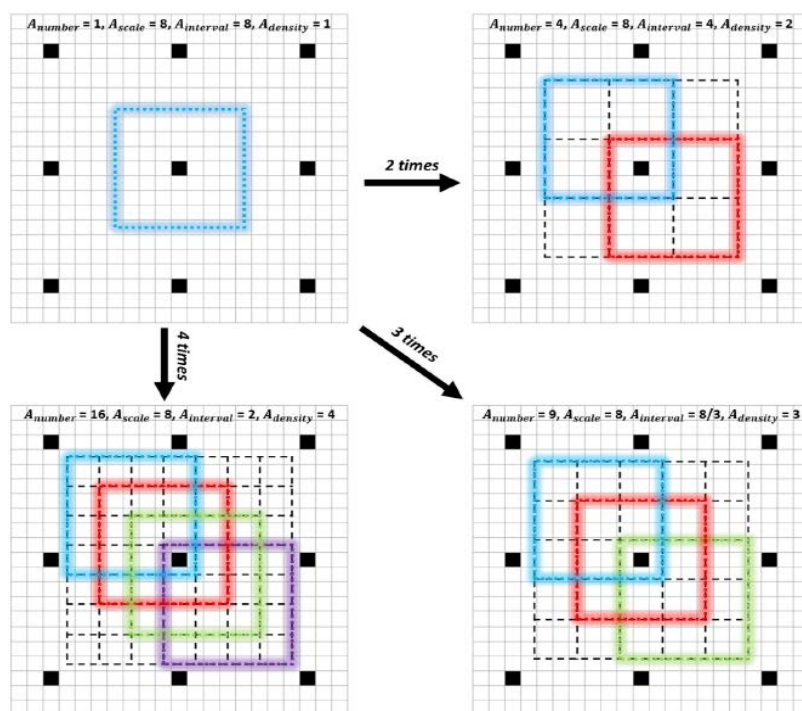


图 5. 锚点框密集化操作

2. 锚点框的关系学习

a) 分块聚合行人检测算法 OR-CNN

基于锚点框的物体检测算法并没有把锚点框之间的关系、物体本身的结构关系、物体之间的上下文关系利用起来,导致在检测遮挡物体时非常容易出现漏检或虚检。而遮挡问题在行人检测这一特定任务中最为显著,行人检测中存在着两种遮挡,一种是类间互遮挡,即行人被其他物体遮住,另一种是类内自遮挡,即行人之间相互遮挡。

行人检测中的类间互遮挡会使行人被其他物体遮住,导致行人的部分特征丢失,并且引入背景中的噪音特征。为了解决这一问题,我们利用行人本身的结构关系,提出了分块遮挡感知的特征融合操作(如图6所示)。该操作先根据经验把行人分成固定的五个部件,再对每一个部件进行遮挡与否的预测。在特征融合的时候,被遮挡的部件赋予一个较低的权重,而可见的部件赋予一个较高的权重,使得融合之后的特征集中于行人可见部分的特征,忽略被遮挡部分的特征。该操作很好地解决了基于锚点框的行人检测算法中类间互遮挡的问题。

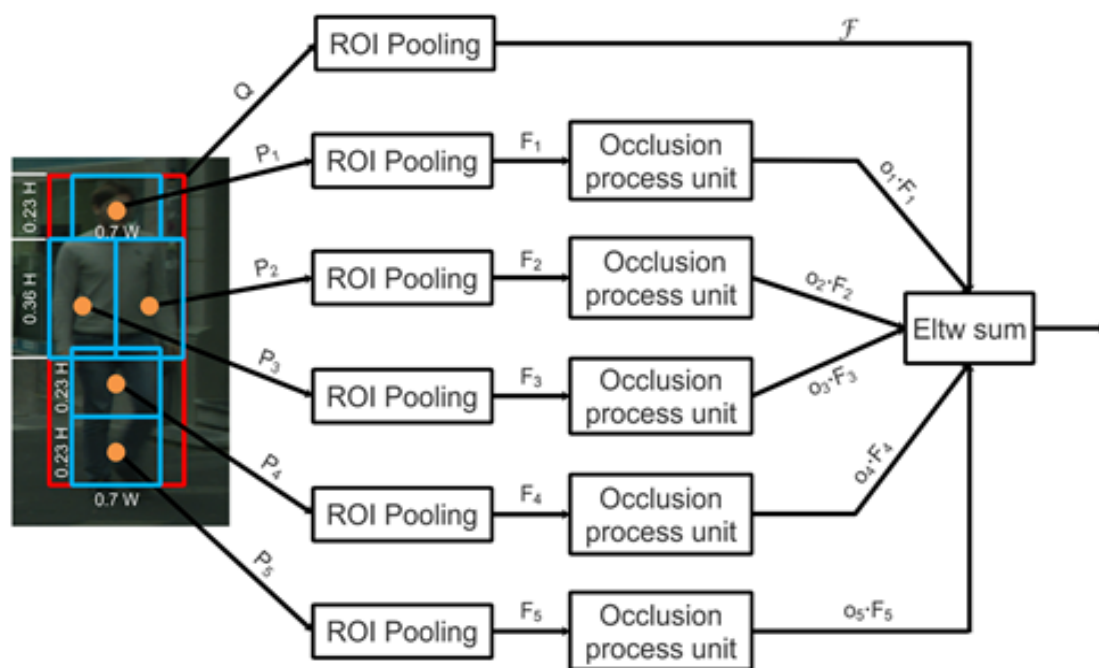
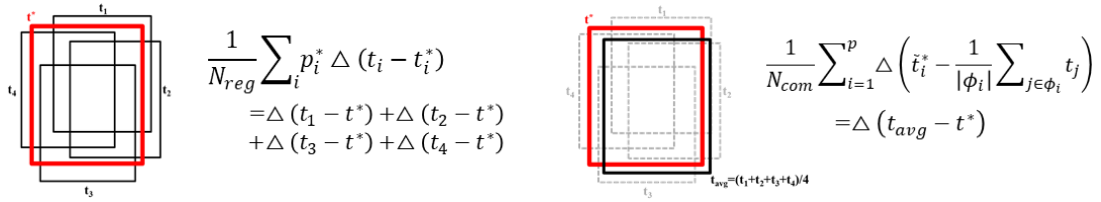


图 6. 分块遮挡感知

行人检测中的类内自遮挡会使得处于两个行人之间的候选区域难以定位造成虚检，或使得两个行人难以区分开来造成漏检。针对这一问题，我们利用锚点框之间的关系，提出了聚合损失函数（如图 7 所示），该函数在原有的回归损失函数的基础上，额外添加了一项紧凑项。该项通过让一个物体所匹配上的所有候选区域的平均候选区域接近该物体，从而让候选区域之间互相帮助进行回归。聚合损失函数使得检测结果在整体上会更加紧凑，有效地解决了行人检测中的类内自遮挡问题。



$$\frac{1}{N_{reg}} \sum_i p_i^* \Delta(t_i - t_i^*)$$

$$= \Delta(t_1 - t^*) + \Delta(t_2 - t^*) + \Delta(t_3 - t^*) + \Delta(t_4 - t^*)$$

$$\frac{1}{N_{com}} \sum_{i=1}^p \Delta\left(\tilde{t}_i^* - \frac{1}{|\phi_i|} \sum_{j \in \phi_i} t_j\right)$$

$$= \Delta(t_{avg} - t^*)$$

图 7. 聚合损失函数

该算法提出后，在常用的行人检测数据集上，都取得了最好的检测精度。该研究的相关研究成果已发表在 ECCV 2018。

b) 人头人体联合检测算法 JointDet

在行人检测任务中，当两个人重叠比例比较大时，一个检测结果就会被抑制掉，这是基于锚点框的物体检测算法的一个通病，即无法检测出两个重叠比例较大的物体。这是因为基于锚点框的物体检测算法，在最后都要进行非极大值抑制的后处理，该处理导致了上述漏检问题。此外，在人头检测任务中，基于锚点框的物体检测算法，经常出现类似人头的虚检，例如头发、手部等。为了同时解决行人检测和人头检测中存在着两个问题，我们利用物体之间的上下文关系，提出了人头人体联合检测算法 JointDet，召回被抑制的人体漏检，同时抑制错误的人头虚检。如图 8 所示为 JointDet 算法的检测框架图。该算法基于著名 Faster R-CNN，在 RPN 阶段只生产人头候选区域，在利用如图 9(a)所示的人头人体结构关系，估计出人体候选区域。这是因为人头的比例固定，生产人头候选区域只需要铺设较少的锚点框，而人体的比例变化很大，要生成人头候选区域则需要铺设大量的锚点框。因此 JointDet 利用人头人体的结构关系，铺设较少的锚点框就能同时地生成人头和人体的候选区域

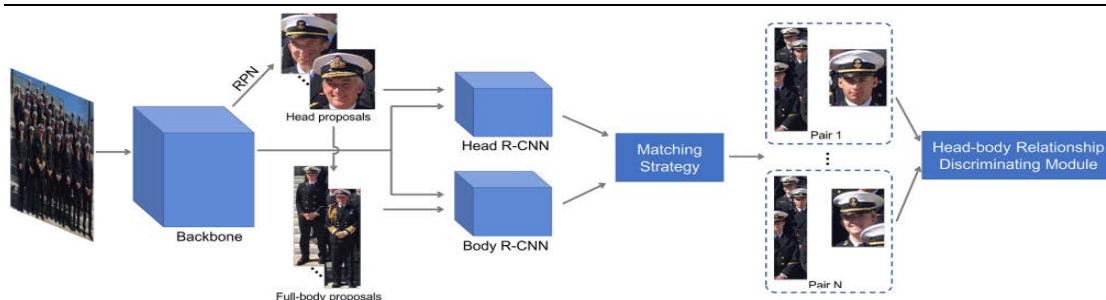


图 8. JointDet 的检测框架

人头人体候选区域生成后，将它们分别送入两个并行的 R-CNN 得到暂时的检测结果。找到未配上对的人头，并从非极大值抑制前找到它对应的人体组成一对，然后学习两者之间的关系。如果符合结构关系，则保留人头检测结果并召回被抑制的人体，如果不符合则抑制掉人头检测见过，具体示例如图 9 所示。该研究的相关研究成果已发表在 AAAI 2020。

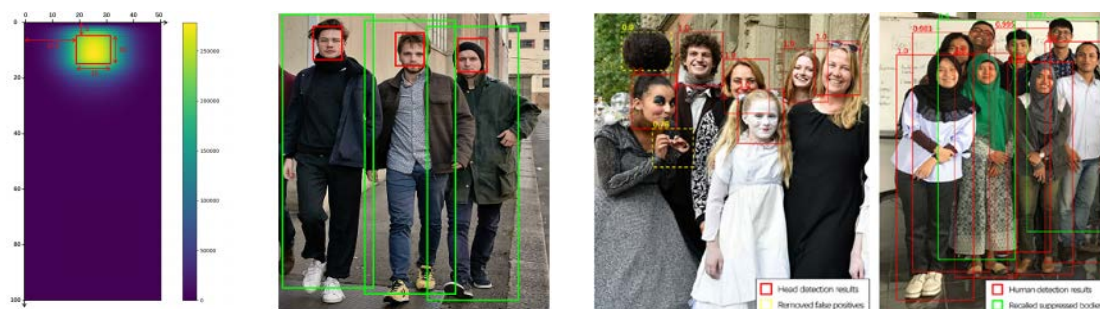


图 9. (a) 人头人体结构关系

(b) 抑制人头虚检和召回人头漏检

3. 锚点框的高效预测

a) 取长补短的物体检测算法 RefineDet

基于锚点框的物体检测算法，如图 10 所示主要分为一步检测法和两步检测法，一步检测法速度更快，而两步检测法精度更高。



图 10. 一步检测法和两步检测法的对比

为了继承两种方法各自的优点,同时克服它们的缺点,我们提出了 RefineDet 物体检测算法。该算法融合了一步检测法和二步检测法的思想,在准确率和速度上达到了很好的均衡。RefineDet 算法是在网络结构上对一步检测法进行改进,以模拟二步检测法的检测过程,从而具备一步检测法的优势,同时能够保持一步检测法的速度。如图 11 所示,该算法由锚点框校正模块和物体检测模块组成,它们之间由传输连接块连接。锚点框校正模块旨在滤除简单的负样本以减少后续分类器的搜索空间,同时粗略地调整锚点框的位置和大小以为后续回归器提供更好的初始化。物体检测模块把经过前者矫正过的锚点框作为输入,以进一步提高回归精度并预测多类别标签。由于锚点框校正模块中的特征被用于二分类,已具备一定的鉴别能力,传输连接块把锚点框校正模块中这些有鉴别力的特征,传递给物体检测模块,以更好地预测物体的位置,大小和类别标签。通过这种网络结构,RefineDet 能够在保持一步检测法的速度前提下,具备二步检测法的二阶段分类、二阶段回归、二阶段特征这三个优点。因此该算法提出时,能够在常用的通用物体检测数据集上,取得最好检测精度的同时,保持着较快的检测速度,在准确率和速度上达到了很好的均衡。该研究的相关研究成果已发表在 CVPR 2018。

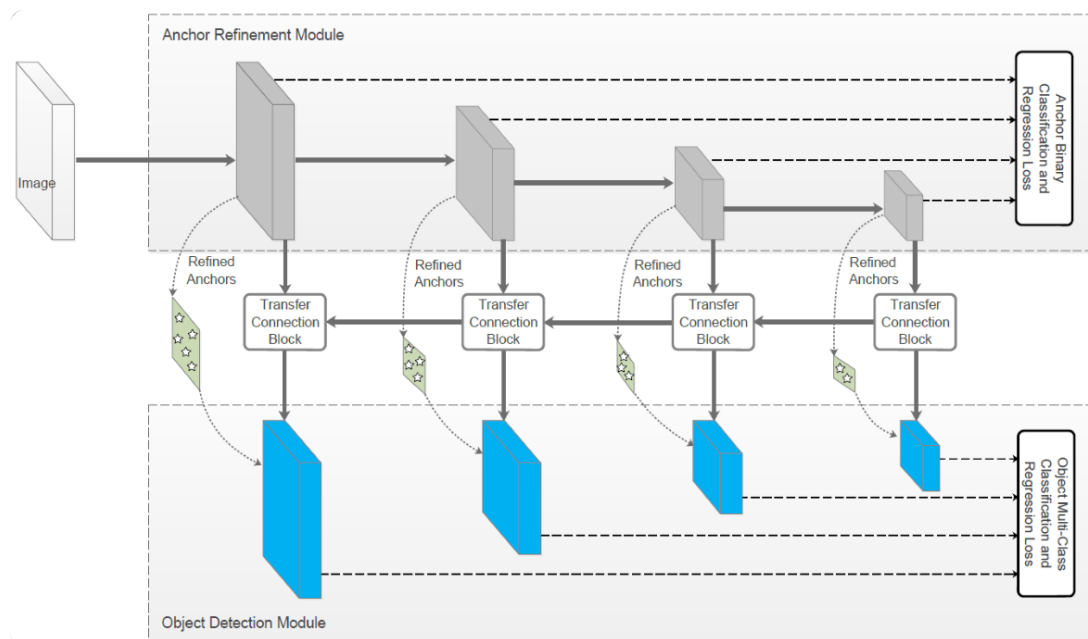


图 11. RefineDet 的检测框架

b) 自适应训练样本选取算法 ATSS

最近无需锚点框的检测算法又流行了起来，主要得益于 FPN 和 Focal Loss 的出现。其中基于中心区域的无锚点框算法就是铺一系列的点而不是框来做检测，这跟基于锚点框的物体检测算法非常类似。两者之间的主要区别有以下两点：

(1) 正负样本的划分，基于锚点框的物体检测算法利用交并比来对锚点框进行正负样本的划分（如图 12(a)所示），而基于中心区域的无锚点框算法利用空间和尺度维度的限制来选取正样本（如图 12(b)所示）；

(2) 回归起始状态，基于锚点框的物体检测算法从框开始回归物体（如图 13(a)所示），而基于中心区域的无锚点框算法则从点开始回归物体（如图 13(b)所示）。

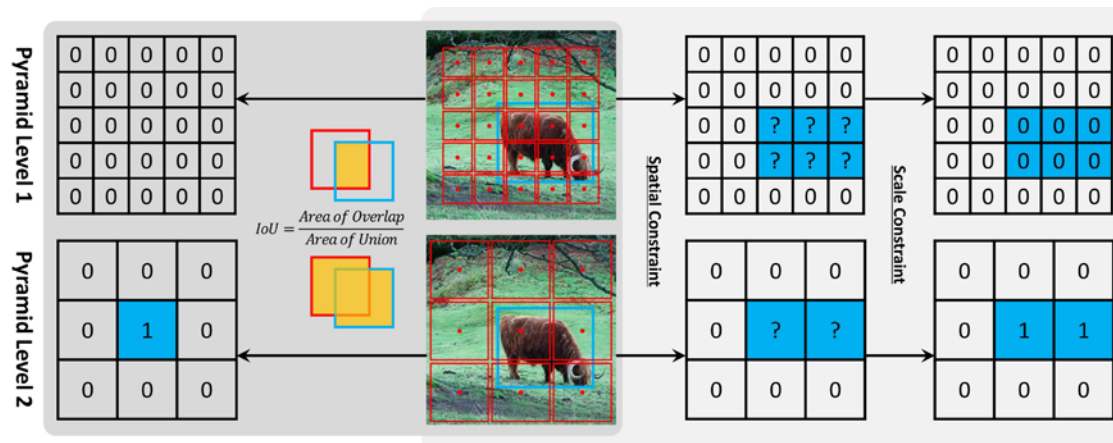


图 12. (a) 基于交并比划分样本

(b) 基于空间尺度限制划分样本

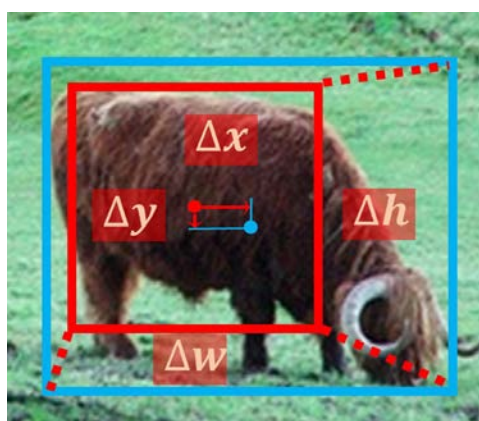
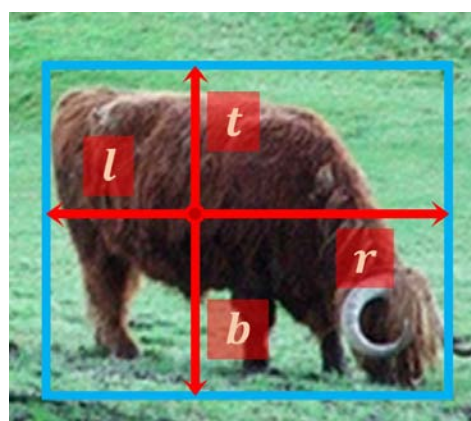


图 13. (a) 从框开始回归



(b) 从点开始回归

经过仔细探索,我们发现无锚点框的物体检测算法和基于锚点框的物体检测算法,两者的本质区别是训练中正负样本的定义不一样,而检测的起点是一个框还是一个点没有影响。鉴于此,我们深入探索了检测中最基本的问题,即如何在训练过程中定义正负样本,提出了一种自适应的正负样本划分方法。具体来说,对于一张输入图片,该算法依次遍历每一个物体,自动地确定它对应的阈值,以用来把锚点框划分为正负样本。对于每一个物体,在每一个特征金字塔上找到距离它中心点最近的 k (默认为 9) 个锚点框。假如特征金字塔有 L 层,则这个物体有 $k*L$ 个锚点框作为正样本的候选。接着统计这 $k*L$ 候选正样本的交叠比的均值 m 和标准差 s ,从而得到该物体划分正负样本的阈值 $t=m+s$ 。候选正样本中,那些交叠比的值比此阈值大的则为最终的正样本。把阈值设置为每个物体的候选正样本交叠比的均值和标准差之和,是因为均值可以衡量锚点框对这个物体的适配程度。如图 14 (a) 所示,均值大表示该物体能匹配到较好的锚点框,因此阈值可以高一些。如图 14 (b) 所示,均值小表示该物体没有较好的锚点框可以匹配,因此阈值要低一些。而标准差是衡量哪一些层适合检测该物体。如图 14 (a) 所示,标准差大表示有一个最合适的层来检测该物体,因此阈值要高一些仅从这个层中选择正样本。如图 14 (b) 所示,标准差小表示有几个层适合来检测该物体,因此阈值要低一些来从这些层中选择正样本。

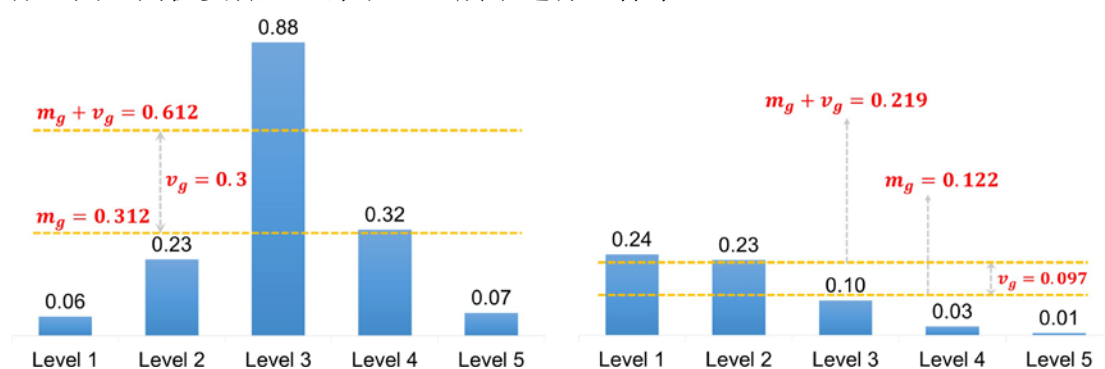


图 14. (a) 高均值和高标准差

(b) 低均值和低标准差

该方法不需要任何超参数,并在不增加任何开销的情况下,极大地提高了检测性能。该研究的相关研究成果已投稿于 CVPR 2020。

4. 学位论文工作进度安排

学位论文涉及到的研究工作基本已经完成,相应成果已经在国际期刊或会议中发表,接下来会继续深入研究一段时间物体检测中锚点框的高效预测这一问题,然后开始撰写学位论文。工作进度安排如下:

2019 年 12 月—2019 年 01 月: 研究检测中正负样本划分的问题,使得基于锚点框的物体检测算法预测的更加高效,并撰写相关期刊或会议论文。

2019 年 02 月—2019 年 04 月: 整理相关科研成果,根据总结的大纲,撰写博士学位论文。

2019 年 05 月—2019 年 06 月: 根据博士学位论文,准备对应的答辩 PPT,进行毕业答辩。

5. 已取得的阶段性成果

目前已发表 23 篇论文,其中第一作者论文 16 篇,包括 11 篇 CCF-A 会议,3 篇 CCF-C 会议,4 篇 SCI 期刊;授权专利 2 项,在申专利 3 项;获得 CCF-CV 学术新锐奖、百度奖学金、国家奖学金、唐立新奖学金、必和必拓奖学金、攀登一等奖学金、戴汝为奖学金、国际人脸检测竞赛冠军、最佳学生论文奖、Valse2018 年度最佳学生论文提名奖等荣誉。

- **Shifeng Zhang**, Longyin Wen, Hailin Shi, Zhen Lei, Siwei Lyu, Stan Z. Li, “Single-Shot Scale-Aware Network for Real-Time Face Detection”, **IJCV**
- **Shifeng Zhang**, Yiliang Xie, Jun Wan, Hansheng Xia, Stan Z. Li, Guodong Guo, “WiderPerson: A Diverse Dataset for Dense Pedestrian Detection in the Wild”, **TMM**
- **Shifeng Zhang**, Xiaobo Wang, Zhen Lei, Stan Z. Li, “FaceBoxes: A CPU Real-Time and Accurate Unconstrained Face Detector”, **Neurocomputing**
- **Shifeng Zhang**, Xiangyu Zhu, Zhen Lei, Hailin Shi, Xiaobo Wang, Stan

- Z. Li, “Detecting Face with Densely Connected Face Proposal Network”, **Neurocomputing**
- **Shifeng Zhang**, Xiaobo Wang, Ajian Liu, Chenxu Zhao, Jun Wan, Sergio Escalera, Hailin Shi, Zezheng Wang, Stan Z. Li, “CASIA-SURF: A Dataset and Benchmark for Large-scale Multi-modal Face Anti-spoofing”, **CVPR**, 2019
 - **Shifeng Zhang**, Longyin Wen, Xiao Bian, Zhen Lei, Stan Z. Li, “Occlusion-aware R-CNN: Detecting Pedestrians in a Crowd”, **ECCV**, 2018
 - **Shifeng Zhang**, Longyin Wen, Xiao Bian, Zhen Lei, Stan Z. Li, “Single-Shot Refinement Neural Network for Object Detection”, **CVPR**, 2018
 - **Shifeng Zhang**, Xiangyu Zhu, Zhen Lei, Hailin Shi, Xiaobo Wang, Stan Z. Li, “S3FD: Single Shot Scale-invariant Face Detector”, **ICCV**, 2017
 - **Shifeng Zhang**, Xiangyu Zhu, Zhen Lei, Hailin Shi, Xiaobo Wang, Stan Z. Li, “FaceBoxes: A CPU Real-time Face Detector with High Accuracy”, **IJCB**, Spotlight, 2017
 - **Shifeng Zhang**, Xiangyu Zhu, Zhen Lei, Hailin Shi, Xiaobo Wang, Stan Z. Li, “Detecting Face with Densely Connected Face Proposal Network”, **CCBR**, Best Student Paper, 2017
 - Cheng Chi*, **Shifeng Zhang**^{*†}, Junliang Xing, Hongwei Qin, Junjie Yan, Zhen Lei, Xudong Zou, “PedHunter: Occlusion Robust Pedestrian Detector in Crowded Scenes”, **AAAI**, 2020. (*共一, † 通信)
 - Cheng Chi*, **Shifeng Zhang**^{*†}, Junliang Xing, Zhen Lei, Stan Z. Li, Xudong Zou, “Relational Learning for Joint Head and Human Detection”, **AAAI**, 2020. (*共一, † 通信)
-

- Xiaobo Wang*, **Shifeng Zhang***, Shuo Wang, Tianyu Fu, Hailin Shi, Tao Mei, “Mis-classified Vector Guided Softmax Loss for Face Recognition”, **AAAI, Oral**, 2020. (*共一)
- Rui Zhu*, Shifeng Zhang*, Xiaobo Wang, Longyin Wen, Hailin Shi, Liefeng Bo, Tao Mei, “ScratchDet: Exploring to Train Single-Shot Object Detectors from Scratch”, **CVPR, Oral**, 2019. (*共一)
- Cheng Chi*, Shifeng Zhang*†, Junliang Xing, Zhen Lei, Stan Z. Li, Xudong Zou, “Selective Refinement Network for High Performance Face Detection”, **AAAI, Oral**, 2019. (*共一, † 通信)
- Xiaobo Wang*, Shifeng Zhang*, Zhen Lei, Si Liu, Xiaojie Guo, Stan Z. Li, “Diversity-induced Ensemble Softmax Loss for Deep Visual Classification”, **IJCAI**, 2018. (* equal contribution)
- Chubin Zhuang, **Shifeng Zhang**, Zhen Lei, Xiangyu Zhu, Jinqiao Wang, “FLDet: A CPU Real-time Joint Face and Landmark Detector”, **ICB**, 2019
- Haibo Jin, **Shifeng Zhang**, Xiangyu Zhu, Yinhang Tang, Zhen Lei, “Learning Lightweight Face Detector with Knowledge Distillation”, **ICB**, 2019
- Chubin Zhuang, **Shifeng Zhang**, Xiangyu Zhu, Zhen Lei, Stan Z. Li, “Single Shot Attention-Based Face Detector ”, **CCBR**, 2018
- Yongming Zhang, **Shifeng Zhang**, Chubin Zhuang, Zhen Lei, “Feature Enhancement for Joint Human and Head Detection ”, **CCBR**, 2019
- Yang Yang, Zhen Lei, **Shifeng Zhang**, Hailin Shi, Stan Z. Li, “Metric Embedded Discriminative Vocabulary Learning for High-Level Person Representation”, **AAAI, Oral**, 2016
- Lu Zhang, Zhi-Yong Liu, **Shifeng Zhang**, Xu Yang, Hong Qiao, Kaizhu Huang, Amir Hussain, “Cross-Modality Interactive Attention Network for

Multispectral Pedestrian Detection”, Information Fusion

- Xuxin Lin, Jun Wan, Yiliang Xie, **Shifeng Zhang**, Chi Lin, Yanyan Liang, Guodong Guo, Stan Z. Li, “Task-oriented Feature-fused Network with Multi-variate Dataset for Joint Face Analysis”, TCYB
- 授权专利 201710379478.3: 人脸检测方法及其装置、计算机可读存储介质、设备; 雷震, 朱翔昱, **张士峰**, 李子青
- 授权专利 201710541087.7: 人脸检测方法及其装置、计算机可读存储介质、设备; **张士峰**, 雷震, 朱翔昱, 李子青
- 在审专利 201810393658.1: 一种基于分块遮挡感知的行人检测方法; 雷震, **张士峰**, 庄楚斌
- 在审专利 201811222696.7: 一种无需预训练网络的通用物体检测算法; 朱睿, 石海林, **张士峰**, 王晓波, 梅涛
- 在审专利 201910502740.8: 单步人脸检测器优化系统、方法、装置; 雷震, **张士峰**, 张永明

6. 课程主要完成情况

已修课程 22 门, 共获得 39 学分, 其中学位课 18 门, 共 35 学分, 具体如下:

类别	公共必修课程及学分	公共选修课程及学分	专业学位课学分要求	总学分要求
学习要求	人文系列讲座 (1 学分) 中国特色社会主义理论与实践研究 (1 学分) 自然辩证法概论 (1 学分) 硕士学位英语 (英语 A) (3 学分)	>=2 学分	>=12 学分	>=30 学分
选课情况	人文系列讲座 (1 学分) 自然辩证法概论 (1 学分) 硕士学位英语 (英语 A) (3 学分) 中国特色社会主义理论与实践研究 (1 学分)	2.0 学分	22.0 学分	35.0 学分
获取学分	人文系列讲座 (1 学分) 自然辩证法概论 (1 学分) 硕士学位英语 (英语 A) (3 学分) 中国特色社会主义理论与实践研究 (1 学分)	2.0 学分	22.0 学分	35.0 学分

各门课程的具体成绩如下：

学年学期	课程名称	分数	学分	学位课
2015-2016 学年秋季学期	模式识别与机器学习	89	3.0	是
	模式识别	94	3.0	是
	人工智能理论与实践	92	3.0	是
	图像处理与分析	95	3.0	是
	矩阵分析与应用	90	2.0	是
	机器视觉及其应用	93	1.0	是
	个性发展与人际交往心理学	93	1.0	否
	人文系列讲座	通过	1.0	是
	自然辩证法概论	82	1.0	是
	博士学位英语	73	2.0	是
	硕士学位英语（免修）	76	3.0	是
	论文写作-2 班	91	1.0	否
2015-2016 学年春季学期	数据挖掘	90	2.0	是
	生物特征识别	96	2.0	是
	视频处理与分析	84	2.0	是
	模式识别研讨与实践	88	1.0	是
	中国马克思主义与当代	80	1.0	是
	中国特色社会主义理论与实践研究	86	1.0	是
2015-2016 学年夏季学期	机器学习与图像/视频分析	60	1.0	否
	统计机器学习	91	1.0	否
2016-2017 学年夏季学期	模式识别与机器学习	80	2.0	是
	最优化算法理论与应用	89	2.0	是

7. 其他

将博士论文题目从“基于锚点预测的物体检测方法研究”变为“物体检测中的锚点框研究”，进一步明确了研究的范围。

主要参考文献

- [1] Viola, P. and M.J. Jones, Robust real-time face detection[J]. International journal of computer vision, 2004. 57(2): p. 137-154.
- [2] Dalal N, Triggs B. Histograms of oriented gradients for human detection[C]//Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on. IEEE, 2005, 1: 886-893.
- [3] Dollár, P., et al., Integral channel features[J]. 2009.
- [4] Dollár, P., et al., Fast feature pyramids for object detection[J]. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2014. 36(8): p. 1532-1545.
- [5] Zhang S, Bauckhage C, Cremers A B. Informed haar-like features improve pedestrian detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2014: 947-954.
- [6] Zhang S, Benenson R, Schiele B. Filtered channel features for pedestrian detection[C]//CVPR. 2015, 1(2): 4.
- [7] Bourdev L, Brandt J. Robust object detection via soft cascade[C]//Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on. IEEE, 2005, 2: 236-243.
- [8] Zhang C, Viola P A. Multiple-instance pruning for learning efficient cascade detectors[C]//Advances in neural information processing systems. 2008: 1681-1688.
- [9] Dollár P, Appel R, Kienzle W. Crosstalk cascades for frame-rate pedestrian detection[M]//Computer Vision–ECCV 2012. Springer, Berlin, Heidelberg, 2012: 645-659.
- [10] Hearst M A, Dumais S T, Osuna E, et al. Support vector machines[J]. IEEE Intelligent Systems and their applications, 1998, 13(4): 18-28.
- [11] Kohavi R. Scaling up the accuracy of Naive-Bayes classifiers: a decision-tree hybrid[C]//KDD. 1996, 96: 202-207.
- [12] Liaw A, Wiener M. Classification and regression by random forest[J]. R news, 2002, 18-22.

- [13] Viola P, Jones M J. Robust real-time face detection[J]. International journal of computer vision, 2004, 57(2): 137-154.
- [14] Dalal N, Triggs B. Histograms of oriented gradients for human detection[C]//Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on. IEEE, 2005, 1: 886-893.
- [15] Felzenszwalb P, McAllester D, Ramanan D. A discriminatively trained, multiscale, deformable part model[C]//Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on. IEEE, 2008: 1-8.
- [16] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[C]//Advances in neural information processing systems. 2012: 1097-1105.
- [17] Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2014: 580-587.
- [18] Uijlings, J.R., et al., Selective search for object recognition[J]. International journal of computer vision, 2013. 104(2): p. 154-171.
- [19] Zitnick C L, Dollár P. Edge boxes: Locating object proposals from edges[C]//European conference on computer vision. Springer, Cham, 2014: 391-405.
- [20] He K, Zhang X, Ren S, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition[C]//European conference on computer vision. Springer, Cham, 2014: 346-361.
- [21] Girshick R. Fast r-cnn[C]//Proceedings of the IEEE international conference on computer vision. 2015: 1440-1448.
- [22] Ren S, He K, Girshick R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks[C]//Advances in neural information processing systems. 2015: 91-99.
- [23] Dai J, Li Y, He K, et al. R-fcn: Object detection via region-based fully convolutional networks[C]//Advances in neural information processing systems. 2016: 379-387.
- [24] Lin T Y, Dollár P, Girshick R B, et al. Feature Pyramid Networks for Object Detection[C]//CVPR. 2017, 1(2): 4.
- [25] Everingham M, Van Gool L, Williams C K I, et al. The pascal visual object classes (voc) challenge[J]. International journal of computer vision, 2010, 88(2): 303-338.
- [26] Lin T Y, Maire M, Belongie S, et al. Microsoft coco: Common objects in context[C]//European conference on computer vision. Springer, Cham, 2014: 740-755.

- [27] Zhu Y, Zhao C, Wang J, et al. Couplenet: Coupling global structure with local parts for object detection[C]//Proc. of Int'l Conf. on Computer Vision (ICCV). 2017, 2.
- [28] Shrivastava A, Gupta A, Girshick R. Training region-based object detectors with online hard example mining[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 761-769.
- [29] Dai J, Qi H, Xiong Y, et al. Deformable convolutional networks[J]. CoRR, abs/1703.06211, 2017, 1(2): 3.
- [30] Kong T, Yao A, Chen Y, et al. Hypernet: Towards accurate region proposal generation and joint object detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 845-853.
- [31] Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, real-time object detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 779-788.
- [32] Liu W, Anguelov D, Erhan D, et al. Ssd: Single shot multibox detector[C]//European conference on computer vision. Springer, Cham, 2016: 21-37.
- [33] Lin T Y, Goyal P, Girshick R, et al. Focal loss for dense object detection[J]. IEEE transactions on pattern analysis and machine intelligence, 2018.
- [34] Liu S, Huang D, Wang Y. Receptive Field Block Net for Accurate and Fast Object Detection[C]// European conference on computer vision, 2018.
- [35] Zhang S, Wen L, Bian X, et al. Single-shot refinement neural network for object detection[C]//IEEE CVPR. 2018.
- [36] Jain V, Learned-Miller E. Fddb: A benchmark for face detection in unconstrained settings[R]. Technical Report UM-CS-2010-009, University of Massachusetts, Amherst, 2010.
- [37] Yang S, Luo P, Loy C C, et al. Wider face: A face detection benchmark[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 5525-5533.
- [38] Dollar P, Wojek C, Schiele B, et al. Pedestrian detection: An evaluation of the state of the art[J]. IEEE transactions on pattern analysis and machine intelligence, 2012, 34(4): 743-761.
- [39] Zhang S, Benenson R, Schiele B. Citypersons: A diverse dataset for pedestrian detection[C]//The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017, 1(2): 3.
- [40] Russakovsky O, Deng J, Su H, et al. Imagenet large scale visual recognition challenge[J]. International Journal of Computer Vision, 2015, 115(3): 211-252.