中图分类号: 学校代码: 10055
UDC: 密级: 公开

南開大學

博 士 学 位 论 文

高效能目标检测的空间均衡与紧致表达的研究

Research towards Spatial Equilibrium and Compact Expression for High-Efficacy Object Detection

论文作者 郑兆晖 指导教师 程明明教授
申请学位 工学博士 培养单位 计算机学院
学科专业 计算机科学与技术 研究方向 计算机视觉
答辩委员会主席 张长青教授 评 阅 人 匿名评阅

南开大学研究生院

二〇二五年五月

摘要

目标检测是计算机视觉领域经久不衰的研究课题之一，同时亦是有向目标检测、3D目标检测、目标跟踪、实例分割等多个高层计算机视觉任务的基石。目标检测应用之广泛，涵盖自动驾驶、医学影像分析、遥感图像目标识别、工地安全等许多场景，具有广阔的应用价值。

高效能是目标检测界长期以来的发展目标，但同时也面临着诸多挑战。高效能意味着目标检测算法需要同时具备高效率与高性能。高效率代表着目标检测器需要有着尽可能快的推理速度，模型尽可能地小而紧凑。高性能则代表着目标检测器的检测精度应尽可能高，有着准确的识别物体能力。本文旨在研究目标检测在通往高效能的道路上所面临的挑战，指出了目标检测社区的两大误区，并从空间均衡与紧致表达两个方面来完善现有目标检测器与评估工具，为指明目标检测新的研究方向提供系统分析与方法论。本文主要贡献包含以下三大方面：

1. 本文提出了区域评估，旨在揭示目标检测算法在取得高性能方面所面临的一大阻碍——空间偏差。空间偏差具体表现为目标检测器十分擅长在图像中央区域检测物体，但对于靠近图像边界区域的物体则检测表现不佳。一直以来，学界鲜有对目标检测空间偏差的探讨，同时缺乏对该现象的系统认识、问题建立、以及可行的解决方法。通过一系列启发性实验，本文对空间偏差的存在性、主要来源、可能的相关因素进行了深入的探索。随后，本文首次建立了一个目标检测新研究问题，称为空间失衡问题，该问题旨在追求空间均衡的目标检测。为此，本文进一步提出首个用于解决该问题的方法，称为空间均衡学习，使目标检测算法在图像空间上尽可能表现均衡。
2. 传统的知识蒸馏大多针对分类头与特征层面展开，无法直接应用于定位头，导致定位头蒸馏效率低下。本文提出了定位蒸馏，将自然界中物体的边界所具有的定位模糊性作为一种重要的知识，将之传递给学生模型后，极大提高学生模型的边界框定位能力，使学生模型拥有更加紧致的表达。同时，本文进一步设计了选择性区域蒸馏算法，使分类知识与定位知识可

因地制宜地传递，带来更高效的蒸馏效率。此外，本文还给出了定位蒸馏与传统的分类蒸馏之间的理论联系，并探索了目标检测知识蒸馏中 logit模仿与特征模仿的优劣性。

3. 检测头网络在浅层级上过于耗时，严重阻碍了检测器取得高效能。本文提出了 SlimHead，一种新型高效能检测头网络，以提高多层级学习模型的紧致性。通过对检测头网络的性能敏感性分析，本文指出了检测头网络的本质属性在于细化特征与定义解空间。该方法具有四大优势：简单、高效、易于推广、低显卡内存占用。

关键词： 计算机视觉；目标检测；高效能；知识蒸馏；检测头网络；评估指标；空间偏差

Abstract

Object detection is one of the most popular research topics in the field of computer vision. It is also the cornerstone of many high-level computer vision tasks such as oriented object detection, 3D object detection, object tracking, instance segmentation, etc. Object detection has a wide range of applications, covering many scenarios such as autonomous driving, medical image analysis, object recognition in remote sensing imagery, construction site safety, etc. It has broad application value.

High efficacy has long been a development goal in the field of object detection, but it faces many challenges. High efficacy means that the object detector needs to have both high efficiency and high performance. High efficiency means that the object detector inference should be as fast as possible and the model should be as small and compact as possible. High performance means that the detection accuracy should be as high as possible, having accurate object recognition capabilities. This thesis aims to study the challenges faced by object detection towards high efficacy, points out two major misunderstandings in the target detection community, and improves the existing object detectors and evaluation tools from the following two aspects: spatial equilibrium and compact expression, which provides a systematic analysis and methodology to point out a new research direction for object detection. The main contributions of this thesis are three-fold:

1. This thesis proposes zone evaluation to reveal a major obstacle faced by object detection in achieving high efficacy - spatial bias. Spatial bias is manifested in that the object detector is very good at detecting objects in the center zone of the image, but performs poorly for objects near the image border. Until now, there has been little discussion in the academic community on spatial bias in object detection, and there is a lack of systematic understanding of this phenomenon, problem establishment, and feasible solutions. Through a series of heuristic experiments, this thesis conducts an in-depth exploration of the existence, main sources, and possible related factors of spatial bias. Subsequently, this thesis first

establishes a new research problem for object detection, called the spatial disequilibrium problem, which aims to pursue spatially equilbrated object detection. To this end, the thesis further proposes the first approach to solve this problem, called spatial equilibrium learning, so that the object detectors can achieve as much equilbrated as possible in the image space.

2. Traditional knowledge distillation is mostly carried out on the classification head and feature level and cannot be directly applied to the localization head, resulting in low distillation efficiency of localization head. This thesis proposes localization distillation, which takes the localization ambiguity of the object edges in nature as an important knowledge. As the localization knowledge passes to the student model, its localization ability can be significantly improved, so that the student model has a more compact expression. Meanwhile, this thesis further designs a selective region distillation algorithm, so that it can selectively distill the classification and localization knowledge for a certain region, achieving more efficient distillation efficiency. In addition, this thesis also gives the theoretical connection between localization distillation and traditional classification distillation, and explores the advantages and disadvantages of logit mimicking and feature imitation in the distillation of object detection.
3. The detection head network is too time-consuming at shallow levels, which seriously hinders the detector from achieving high efficacy. This thesis proposes SlimHead, a new high-efficacy detection head network, to improve the compactness of the multi-level learning models. Through the performance sensitivity analysis of the detection head network, this thesis points out that the intrinsic properties of the detection head network are to refine features and define the solution space. This proposed method has four major advantages: simple, efficient, easy generalized, and low GPU memory usage.

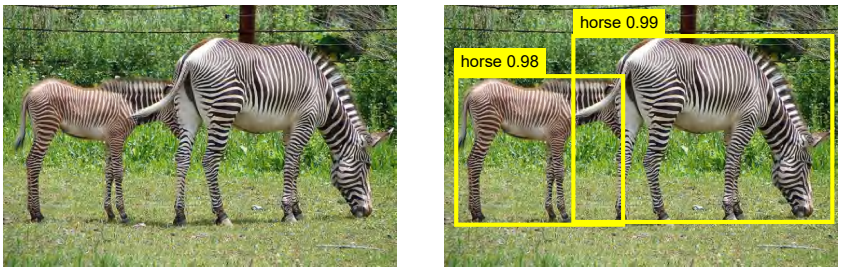
Key Words: Computer Vision, Object Detection, High Efficacy, Knowledge Distillation, Detection Head Network, Evaluation Metrics, Spatial Bias

目录

第一章 绪论

1.1 研究背景与意义

目标检测，如图 1.1所展示，任务定义很简单：识别出给定图像内给定类别的所有物体，并使用矩形框将其标出。作为计算机视觉领域下的经典老牌任务，目标检测在数十年来一直受到了广泛的关注。无数的学者从各个方面，包括检测框架、基础网络、算法优化、数据集、评估方法等，推动着目标检测持续不断地向前发展。其下游任务也衍生出了多个庞大的研究分支，譬如遥感检测 [1]、3D目标检测 [2]、目标跟踪 [3,4]、实例分割 [5]等，一同构建出了高层计算机视觉大厦。



(a) 原图 (b) 目标检测结果图

图 1.1: 目标检测示意图

目标检测在现实世界中有着丰富的应用场景，人们几乎可以随处可见它们的身影。智能支付系统中的人脸识别，道路监控系统中的行人检测，停车场收费系统中的车牌检测，智能机器人抓取中的物体识别，扫地机器人中的障碍物检测，新冠疫情下被各大商场、火车站、机场所部署的口罩检测，智能手机拍照中的人脸跟踪，在线网购平台中的商品检索，目标检测的应用已然无处不在。在当前人工智能的浪潮下，目标检测通常被认为是计算机理解现实世界的重要途径之一。图 1.2展示了几种常见的目标检测在现实中的应用部署与应用场景。

应用部署：智能手机 [6–8]、家用汽车 [9–11]、机器狗 [12,13]、智能摄像头 [14,15]、车牌识别一体机 [16–18]、云端服务器 [19,20]、无人机 [21,22]、无人

战斗机 [23, 24]、扫地机器人 [25–27]、人造卫星 [28, 29]、智能机械臂 [30–32]、配送机器人 [33–35]等。

应用场景：自然场景 [36, 37]、建筑工地 [38–42]、口罩检测 [43–45]、医学影像分析 [46–49]、商品零售 [50, 51]、人脸检测 [52–55]、智慧安防 [56–58]、场景文本识别 [59–62]、道路监控 [63–66]、车牌检测 [67–69]、明火检测 [70, 71]、森林火情预警 [72–75]、无人机航拍 [22, 76]、卫星遥感 [77–80]、自动驾驶 [81, 82]、合成孔径雷达 [83–86]等。

图 1.2: 目标检测应用部署和应用场景示例

除了上述关于目标检测的工业应用以外，目标检测其本身的框架与算法优化也历经了数十年的版本迭代 [87]。学界早期的研究多依赖于人工特征提取法，使用预先手工定义的模板，亦称为卷积核，对图像进行滑动卷积，如 SIFT 特征 [88]、HOG 特征 [89]等方法能够提取到与目标相关的特征。再对所提取到的这些特征使用支持向量机（SVM）进行分类。最后经过非极大值抑制（NMS）算法去冗余，得到最终的检测结果。较为有代表性的成果如 DPM [90]，走组合到整体的特征提取路线，因其优越的检测性能表现，DPM 连续三年蝉联 PASCAL VOC 2007 检测大赛冠军。

2012年，随着 AlexNet [91]的问世，卷积神经网络（Convolutional neural network，CNN）蓬勃发展，向着深度而强大的方向前进。目标检测的发展同样深受深度学习的影响，最具代表性的目标检测流程为 R-CNN算法，首次将CNN应用于目标检测，学界从此开启了深度学习时代。 R-CNN 采取区域提案 → 特征提取 → 分类器 → 非极大值抑制的流程。滑动窗口法，如 OverFeat [92]，是区域提案法的前身，其使用不同大小、比例的边界框在一幅图像上滑动，再对每一次滑动所得区域使用卷积神经网络（CNN）进行特征提取与分类，来得到每一个滑动窗口的分类得分。人工特征提取的方式也就此退出了舞台。然而以穷举法为核心思想的滑动窗口法容易造成大量计算开销，且数以十万计的窗口也包含着大量与目标无关的区域。 R-CNN [93]通过选择性搜索（Selective Search） [94] 生成一组具有代表性的感兴趣区域（Region of Interest，RoI），其目的是在原始图像中找到一系列可能包含目标的候选区域。一般而言，这些RoI的数量约为2000个左右。这不仅减少了需要处理的区域数量，提高了检测效率，而且这些 RoI通常能够较好地覆盖图像中的潜在目标，使得检测准确性也得到改进。同年，MS COCO [37]大规模检测数据集问世，学界由此进入了更具有挑战性的研究场景。 2015年，R-CNN得到进一步改进，学界陆续提出了 Fast R-CNN [95]以及 Faster R-CNN [96]。耗时的选择性搜索被剔除，取而代之的是区域提案网络（Region Proposal Network，RPN），其能够使 RoI 以一种更高效更精确的方式被生成。在 RPN 中，锚框（Anchor Box）的概念被首次提出。锚框是一系列中心点被规则放置，具有多种尺度、宽高比的初始候选框，密集铺设在特征图上。经过第一次正负样本定义，也称为标签分配，一些高质量锚框被选为阳性样本，其余的为阴性样本，RPN 则对这些锚框进行二分类学习，输出约256个 RoI。这其中，包含高质量的阳性 RoI 与部分阴性 RoI，用于 Fast R-CNN 进行第二阶段的分类与坐标回归。由于总计使用了两次标签分配与边界框回归，Faster R-CNN 也由此成为了经典的双阶段目标检测器。

2016年， You Only Look Once （YOLO） [109]横空出世，开启了单阶段目标检测的时代。随后的2017年，特征金字塔网络（FPN） [105] 被提出，将不同大小的物体依照标签分配的结果分至不同的层级上进行学习，大物体由深层学习，小物体由浅层学习，极大降低了目标检测的优化难度。同年，RetinaNet [106]以 FPN 为基础，搭配 focal loss 构建了单阶段目标检测模型。此时的单阶段目标检测器，已经可简单地视为多分类任务下的 RPN。

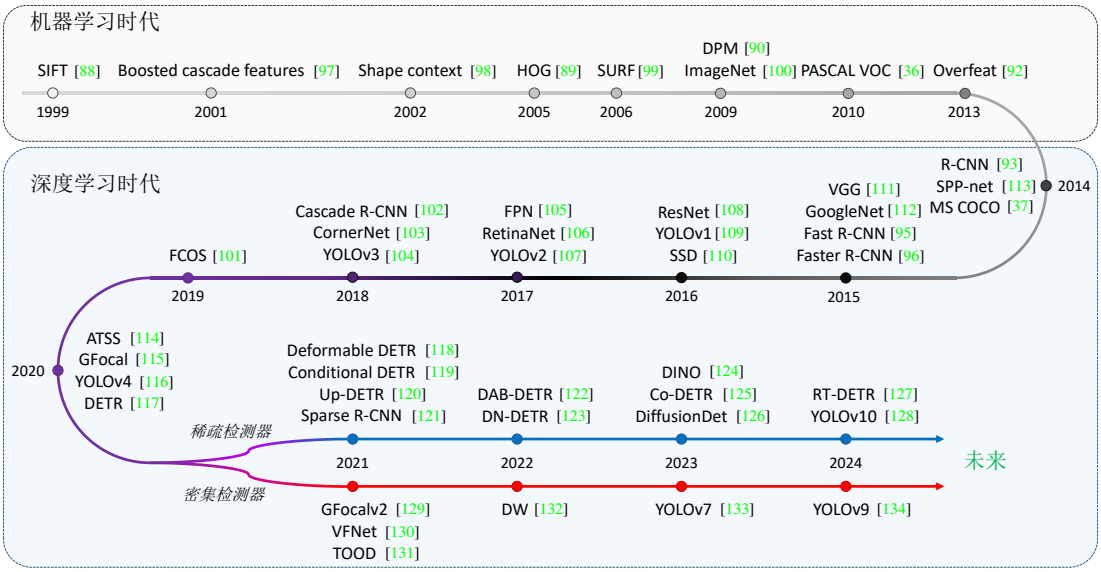


图 1.3: 目标检测发展时间线

YOLOv2 [107]，YOLOv3 [104]都分别注意到了标签分配在目标检测精度优化中的重要性，标签分配算法也引起了学界新一轮的研究热潮。2019年，FCOS [101]在 RetinaNet 的基础上，更改了边界框表示形式，改为预测采样点到真实框上、下、左、右四条边距离（t,b,l,r）。得益于这一边界框表示，FCOS使用中心采样策略，将标签分配的结果限制在真实框内部的中心一小部分区域中。同时由于 FCOS 不再需要锚框的参与，也因此被称为基于回归的无锚框目标检测器（regression-based anchor-free）。然而，基于回归的有锚框目标检测器与无锚框目标检测器究竟孰优孰劣？ ATSS [114]通过严格的控制变量，探究了这两种不同类型的目标检测器的本质差异正是在于标签分配。在二者使用相同的标签分配算法时，基于回归的有锚框与无锚框目标检测器在性能表现上并无本质差异。密集回归目标检测器的标签分配算法也从静态算法逐渐向动态算法发展，后续逐渐开发出了 FreeAnchor [135]，AutoAssign [136]，PAA [137]，OTA [138]，DW [132]等。

2020年，Transformer模型 [139]被首次引入目标检测领域。由 Carion 等人开发的 DETR [117]目标检测器，在 CNN 主干网络后接 Transformer 编码器与解码器，以集合预测（set predicttion）的方式输出固定数量100个的边界框查询（Query）。标签分配则以匈牙利算法 [140]二分图匹配来完成。由于标签分配以一对一的方式完成正样本与真实框的匹配，因而 DETR目标检测器可以输出稀疏的检测框，无需 NMS 算法去除冗余框，从而完成端到端的目标检测。然

而，也正因为一对一匈牙利匹配算法的低效性，以及 Transformer模型缺乏对局部位置的感知能力，DETR模型需要300轮的训练才能达到与 Faster R-CNN相当的精度水平。因此在后续的改进方案中，学界再次提出了类似锚框的机制来引入位置先验信息以加快模型收敛速度，如 Conditional DETR [119]，DAB-DETR [122]。还有如 Deformable DETR [118]引入了多层级预测的方式，结合可形变注意力机制，使得 DETR收敛速度极大加快。 DN-DETR [123]引入了查询去噪操作，其向 Transformer 解码器输入带有噪声的真实框,并训练模型重构原始真实框，有效降低了二分图匹配的优化难度。越来越多的研究者注意到了密集回归目标检测器在标签分配算法的优越性，从而提出结合一对多匹配的方式来辅助模型收敛。这些算法在训练过程中结合了密集回归目标检测的一对多标签分配算法，同时又保留原始的一对一匈牙利匹配。在推理阶段，那些一对多匹配的分支将被去掉，从而可以不增加模型的推理开销完成端到端目标检测。这类方法的代表性成果有 Group DETR [141]、H-DETR [142]、Co-DETR [125]等。

综上所述，目标检测不仅在计算机视觉与模式识别领域有着极高的学术研究价值，同时还在辅助医疗、工地安全、自动驾驶、国防建设、智慧城市交通等领域有着广阔而重要的工业部署与应用价值。本文将深入研究目标检测技术，探索该领域所存在的问题，提出有效的解决方案，从而将目标检测算法向着更快更强的方向发展，以期望在现实复杂场景中展现出更高效能的应用表现。

1.2 困难与挑战

尽管目标检测技术的发展已有数十年的历史，然而学界对目标检测仍存有两大误区：一是过于依赖当前的权威评价指标平均精度（AP），并以极致追求 AP 指引目标检测的发展，而 AP 却无法衡量目标检测所面临的空间偏差问题；二是学界长期以来忽略了定位头知识蒸馏，这是因为传统分类蒸馏与特征模仿无法直接用于定位头。以上两个误区直接导致了学界对目标检测模型的空间均衡性的关注缺失，以及对紧致表达的不充分改进。以下总结了目标检测算法在走向高效能的道路上所面临的三大具体困难与挑战。

- 由于目标检测的应用场景往往较为复杂，且数据服从严重的摄像师偏差，这可能导致通用目标检测器在图像空间域上表现出不均衡的检测能力，即空间偏差。现有的评价指标通常以图像级为单位，在全图范围内评估检测器的检测能力，难以捕捉检测器存在的空间偏差。如何开发客观的、细粒

度更强的评价标准成为当前通用目标检测界亟待解决的难题。此外，当前学界对目标检测的研究鲜有对空间偏差的深入探索，人们对空间偏差的成因、主要来源、影响因素尚没有明确认识。空间偏差的研究空白，阻碍了目标检测研究的进一步发展。此外，目前学界尚没有形成对空间失衡问题的建立，那么自然而然也就缺乏针对该问题的解决方法的提出，这将进一步严重影响目标检测在现实应用中的空间鲁棒性。

- 以卷积神经网络为基础的通用目标检测器通常需要处理大量的图像数据，并且模型结构复杂，这导致了推理过程中需要消耗大量的计算资源，不利于低端边缘设备的部署与实时检测。这不仅增加了硬件成本，还可能导致推理时间较长，降低了模型的实用性。知识蒸馏作为模型压缩的一种经典算法，对于学习轻量级目标检测器至关重要。然而现有的目标检测知识蒸馏却受限于定位分支蒸馏效率低下的问题，如何得到紧凑而强大的目标检测器，并且保证模型的高效率与高精度是一大挑战。
- 由于目标检测模型的检测头网络大多采用多层级学习的方式来预测各种尺度的物体，负责小物体的浅层级特征图分辨率较大，往往占据大量计算量，成为了高效能目标检测的一大瓶颈。目前学界鲜有对该问题的探索。更多计算单元的堆叠反而进一步加重了多层级学习的计算负担。如何得到一个更加紧凑的多层级学习框架，将是走向高效能目标检测的一大挑战。

1.3 研究内容及组织架构

鉴于目标检测巨大的应用前景，考虑现存的困难与挑战，本文以通用目标检测为研究重点，从目标检测知识蒸馏、检测头网络、评估这三个方面来扩展和完善现有的研究。图 1.4梳理了本文核心内容之间的关系。围绕上述研究内容，本文共分为六章，具体组织如下：

- 第一章主要介绍目标检测的研究背景与意义、所面临的困难与挑战、本文的研究内容及组织架构。
- 第二章先回顾了包括目标检测基础模型架构、边界框表示与优化、目标检测知识蒸馏、鲁棒性研究、多层级学习等研究现状，最后列举了常用的目标检测数据集与评估方法。
- 第三章，定位蒸馏：高效能知识蒸馏技术。该章节将提出一种新的目标检测蒸馏技术，称为定位蒸馏。其主要思想基于边界框的概率分布表示，将

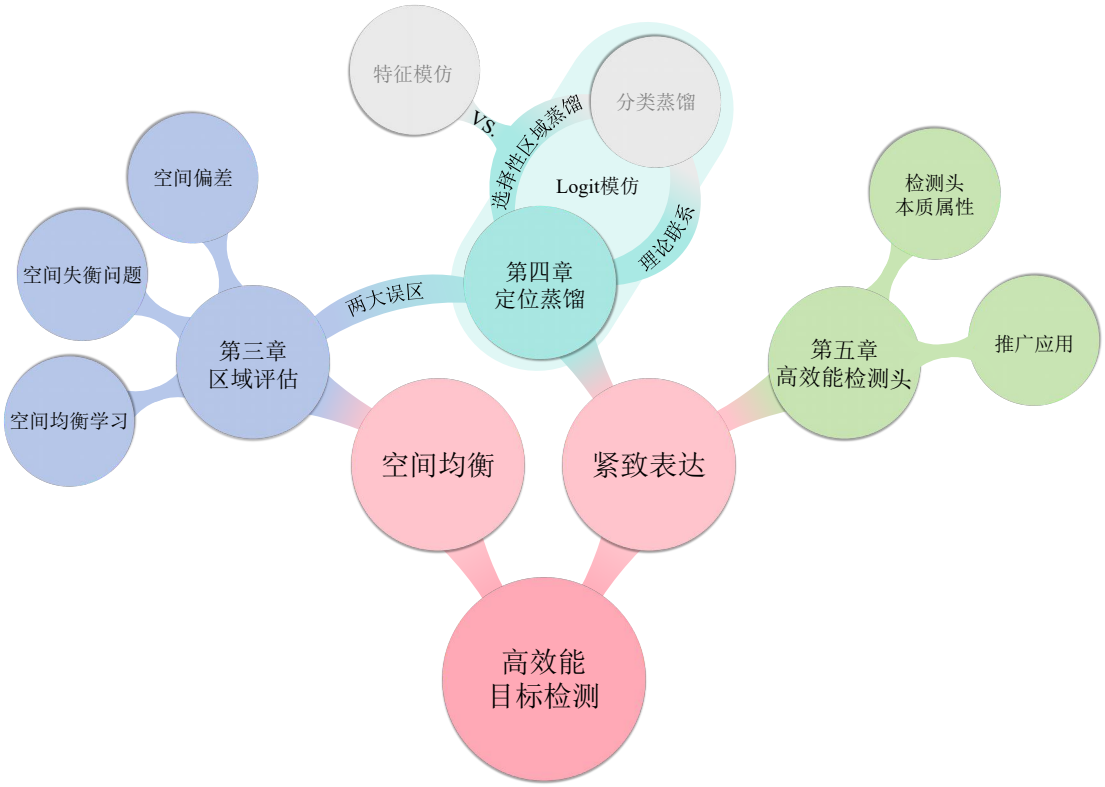


图 1.4: 本文主要章节贡献

自然界中物体的边界所具有的定位模糊性作为一种重要的知识，将之传递给学生模型，以提高学生模型的边界框定位能力。同时，本文进一步设计了选择性区域蒸馏算法，使分类知识与定位知识可因地制宜地传递，带来更高效的蒸馏效率。该章节还将讨论定位蒸馏的性质及其优化特性，讨论知识蒸馏中的两大技术“logit模仿”与“特征模仿”的优劣性，给出logit模仿与特征模仿在训练过程中的差异以及所学习到的学生检测器的表现特点。通过大量的定量实验分析，本文将揭示两种蒸馏算法的优化特性与适用情况。

- 第四章，SlimHead：高效能检测头网络。该章节将提出一种新型检测头网络，用于得到更加紧凑的多层级学习模型，实现更好的速度-精度权衡。通过对检测头网络进行性能敏感性分析，该章节总结了检测头网络的本质属性在于细化特征与定义解空间。 SlimHead的设计能够综合考虑检测头网络的两个本质属性，从而拥有四大优势：简单、高效、易于推广、降低显卡内存占用。

- 第五章，空间偏差：揭示通往高效能的新阻碍。该章节将提出一种新的目标检测评估方法，称为区域评估。该评估方法将传统的评估标准推广至更加一般的情形，以局部视角对目标检测器进行评估，从而补充了一系列区域评估指标，这些指标能够以更高细粒度来刻画目标检测器在图像空间域上的性能表现。区域评估法首次定量确定了目标检测器空间偏差现象的存在性，并为探讨空间偏差的主要来源与可能的相关因素提供了方法论。更进一步，该章节还将为目标检测领域建立一个全新的研究课题——空间失衡问题。该问题旨在追求空间均衡的目标检测，使目标检测算法在图像空间上尽可能表现均衡。该问题属于目标检测不平衡问题下的一个子问题，却鲜少被关注。随后，该章节将给出空间失衡问题与类别不平衡问题之间的联系，二者共享等价的表达形式。作为对该问题的初步探索，本章节还将提出首个解决方案，空间均衡学习。空间均衡学习主要受到类别不平衡问题的相关解决方案所启发，包括重采样法、代价敏感学习等，可实现更加空间均衡的目标检测。
- 第六章总结了本文的主要研究成果，并对未来的研究提出了展望。

第二章 相关工作

目标检测在通往高效能的道路上尽管阻碍重重，但学界仍有丰富的探索，其主要关注点集中在目标检测鲁棒性研究、加强边界框定位质量，知识蒸馏压缩模型，构建多层级学习基础模型架构、构建数据集等方面。本章节将简要介绍上述相关工作，它们其中一些也将成为本文研究的基石与灵感来源。

2.1 目标检测鲁棒性研究

2.1.1 数据不平衡问题

目标检测器的性能常受到数据不平衡问题的影响，这是困扰学界与工业界的一大难题，由来已久。设 $\{X,G\}=\{x_i,g_i\}_{i=1}^{n}$ 是一个样本-标签对的集合，每一个样本 x_i 都有一个真实标签 g_i 与之对应。模型的训练与优化通常在 $\{X,G\}$ 的一个子集上进行，神经网络在每个训练周期都会完整采样过一遍训练集。那么，数据不平衡问题通常与 $\{X,G\}$ 的固有属性有关。在已有的研究中，数据不平衡问题体现在两个方面，其一是类别不平衡问题，其二是前景背景采样不平衡。

（1）类别不平衡

在该问题下，样本 X 由一系列子集 $X_1,X_2,\cdots,X_c$ 所构成，这些子集是依据类别而划分。例如，一个自然场景下的数据集可能包含多个类别，由人、车、猫、狗等等类别共同组合而成。显然，每一个子集 X_i 的样本数量在 c 个类别中是不平衡的，通常会形成一个明显的长尾分布 [143–146]。那么，类别不平衡问题天然地会导致训练过程中的不平衡采样，使得所学习到的模型对头部类别表现较好，而尾部类别表现较差。针对该问题，一些重采样方法 [147, 148]与代价敏感学习 [149, 150]是类别重平衡的主流解决方案。

（2）前景背景采样不平衡

该问题同样由数据本身所引发。大量的锚点平铺在背景区域上，这些锚点自然被采样为负样本，从而主导了梯度流，使模型的学习偏向于负样本（背景）。在该问题下，X 可被分为 X_{neg} 与 X_{pos} 两个子集，使得 $X=X_{neg}\bigcup X_{pos}$。

负样本集合 X_{neg} 可被视为 X_{pos} 集合的补集，其真实标签是“背景”，没有边界框标签。针对该问题的解决方案也是类似的，如重采样方法 OHEM [151]、Guided Anchoring [152]、IoU-balanced sampling [153]，还有代价敏感学习 Focal loss [106]、GHM loss [154]、以及 PISA [155].

2.1.2 卷积神经网络的鲁棒性研究

平移不变性并不被卷积神经网络（CNN）所完美保持，因为其忽略了经典采样定理，这一观点已在多项研究工作中 [156–160]有过探讨。一个微小的图像变换都可能会导致预测输出发生巨大变化，从而阻碍分类器的鲁棒性。 Zhang R. [157]分析了最大池化算子的缺陷，其提出注入抗锯齿来提高深度网络的鲁棒性。 Lopes等人 [161]通过提出块高斯增强实现了更好的鲁棒性-准确性的权衡。在较长的空间范围内的鲁棒性，最近的研究 [159, 162]表明卷积神经网络可以利用物体的绝对位置作为图像分类的附加信息。 Islam等人 [163]进一步扩展了卷积神经网络是根据通道维度的排序对位置信息进行编码。在 [164]中，研究者提出了一种空间无偏的 StyleGAN2 [165]模型，用于解决人脸数据集 [166]中由于摄影师偏差所导致的图像边界中的人脸生成扭曲问题。 Gergely等人 [167]通过经验发现当平移图像以使物体更靠近图像边界时，分类准确率会下降。 Islam, M. A. 等人 [158]的工作揭示了语义分割存在边界效应，其中车辆分割的质量与区域内车辆密度呈现高度相关性。 Manfredi等人 [168]提出了一种贪婪的 AP变化近似方法，其通过将图像平移几个像素来测量目标检测器的平移等变性，但它不可避免地需要数倍的推理时间才能完成评估。

综上所述，卷积神经网络天然面临着会导致其鲁棒性剧烈下降的问题，这些问题可能来源于某个网络算子或模块的运算属性所决定，也可能是来源于譬如摄像师偏差的数据不平衡问题。然而学界目前仍然缺乏对目标检测的鲁棒性的深入探讨，特别是缺乏一种高效的度量方法能够捕捉目标检测器所存在的鲁棒性缺陷、变化幅度、以及相关因素。

2.1.3 局部视角下的评估

由于卷积神经网络在图像全空间范围内可能出现性能表现上的波动，因而对模型的性能评估也就不能简单地以全图指标概之。目前学界已经广泛证明，全局评估与人类视觉系统（Human Visual System，HVS）并不一致 [169–174]，因为人眼通常容易受局部的表现瑕疵的影响而降低对模型性能的印象。因而

从局部视角进行评估就成为了一种科学的模型评价方法，这在图像质量评估（Image Quality Assessment，IQA）领域 [175]中具有优势。一个全局指标无法反映空间上非平稳的模型能力，已成为共识。这一观点甚至可以追溯到1982年的一项早期研究 [176]，当时研究者提出，如果使用局部测量而不是全局测量，质量测量可能会有所改善。在 IQA中，通常采用两阶段结构。在第一阶段，对图像质量进行局部评估。局部评估过程通常会生成质量图。为了将这些质量图转换为全局评估分数，在 IQA的第二阶段会应用池化算法。例如 Wang等人提出了 Mean-SSIM [169]来获得用于图像失真评估的空间平滑测量，其关键是计算每个滑动窗口的局部 SSIM，然后取平均值。 3-SSIM [177]提出为边缘、纹理和光滑区域分配不同的权重。 Larson等人 [178]引入了可见性加权局部 MSE来确定感知失真，其中图像被分成 16×16 个块，相邻块之间有75%的重叠。 NIQE [179]指标引入了块选择，以关注信息丰富的图像块。 Chen等人 [180]提出使用 Landmark Distance（LMD）来重点衡量模型所生成的唇部运动的质量。 Sun等人 [181]提出了加权球面均匀 PSNR（WS-PSNR）来为不同的像素提供不同的权重。 GMSD [182]利用像素梯度幅度相似性来捕捉图像局部质量。 Fan等人 [183]提出了一种用于显著物体检测的 S-measure，首先将图像分成4个方格，并为每个局部 SSIM分配不同的权重。 Bosse等人 [184]尝试使用基于卷积神经网络的方法来学习局部图像质量。还有一些方法 [173, 185–187]将显著性图纳入 IQA指标，因为显著的位置可以帮助预测人类观察者所感知的图像质量。

综上所述，局部评估有助于描述图像块之间的小细节和结构相似性。尽管局部评估几十年来在许多计算机视觉应用的评估系统中一直很流行，但在对象检测领域尚未得到充分研究。大多数 IQA方法主要用于像素预测任务，例如图像恢复 [188–190]、显著/伪装物体检测 [191–193]和图像超分辨 [194, 195]，它们不能直接应用于实例预测任务，例如通用目标检测。

2.2 边界框的表示与优化

目标检测由两大核心任务构成：分类与定位。一个目标检测器需要识别出图像中给定类别的物体，再用矩形框来将它们标出。由任务定义来看，目标检测器的对物体的定位能力将直接决定该检测器的性能表现。而在评估指标方面，现有公认的较为权威的评估指标平均精度（Average Precision，AP）则同时考虑

了目标检测器的分类能力与定位能力。因此，学界对于目标检测器定位能力的优化也进行了丰富的探索。

2.2.1 边界框表示

一个边界框通常由一个四维坐标表示，在已有的文献 [96, 101]中，两种表示较为流行。一种是预测框到真实框的映射 $\{\delta_x, \delta_y, \delta_w, \delta_h\}$，其编码了预测框中心点 (x, y) 与宽高 (w, h) 到真实框的归一化映射。另一种为 $\{t, b, l, r\}$，则是编码了采样点到真实框的上、下、左、右四条边的距离。根据 ATSS [114]一文，在基于回归的目标检测器中，上述两个边界框表示在检测性能上没有本质差别。由于二者可以互相转化，本文统一使用 $\{t, b, l, r\}$ 表示，并在下文中不再予以区别对待。

在 Softer-NMS [196]与 Gaussian YOLOv3 [197]中，研究者观察到物体的边缘天然存在着定位模糊性，即物体的边界可能与背景环境高度混合，从而影响模型对物体边界的判断。为了捕捉目标检测器的定位不确定度，Softer-NMS [196]与 Gaussian YOLOv3 [197]不约而同考虑了一种基于高斯分布的边界框表示，该表示的核心思想是对每条边使用均值与方差来进行表示，均值即表示一条边的预测位置，而方差则代表该边的定位不确定度。于是，经过有监督学习后，越确定的边界，其方差就越小，而越模糊的边界，方差则相应越大。有了方差作为对定位模糊性的刻画后，Softer-NMS将方差用于后处理算法的坐标加权非极大值抑制中，使得越确定的边界加权权重大，而越模糊的边界加权权重小。而 Gussian YOLOv3则提出利用方差对分类得分进行惩罚，以降低定位较为模糊的预测框的分类得分。

继高斯分布表示后，学界又一次推出了全新的概率分布表示 [115, 198]。从概率分布的角度来看，高斯分布是一种对称的理想化的表示，无法应对复杂的现实场景，因为现实中边界框的定位模糊性可以是任意的。因此，建立一种任意形状的概率分布将更加贴近现实应用场景。 图 2.1展示了边界框表示的发展之路，对边界框的刻画从原先的狄拉克 Delta分布的单一逐渐朝向丰富。

2.2.2 边界框回归

定位任务的核心是使得模型预测框尽可能逼近真实框。边界框回归是实现这一目标的有效途径。设 $\mathcal{B}_p$ 是模型预测框，$\mathcal{B}_g$ 是真实框，边界框回归损失将施加在二者之间以衡量衡量预测框与真实框的误差，模型将依据梯度下降法来

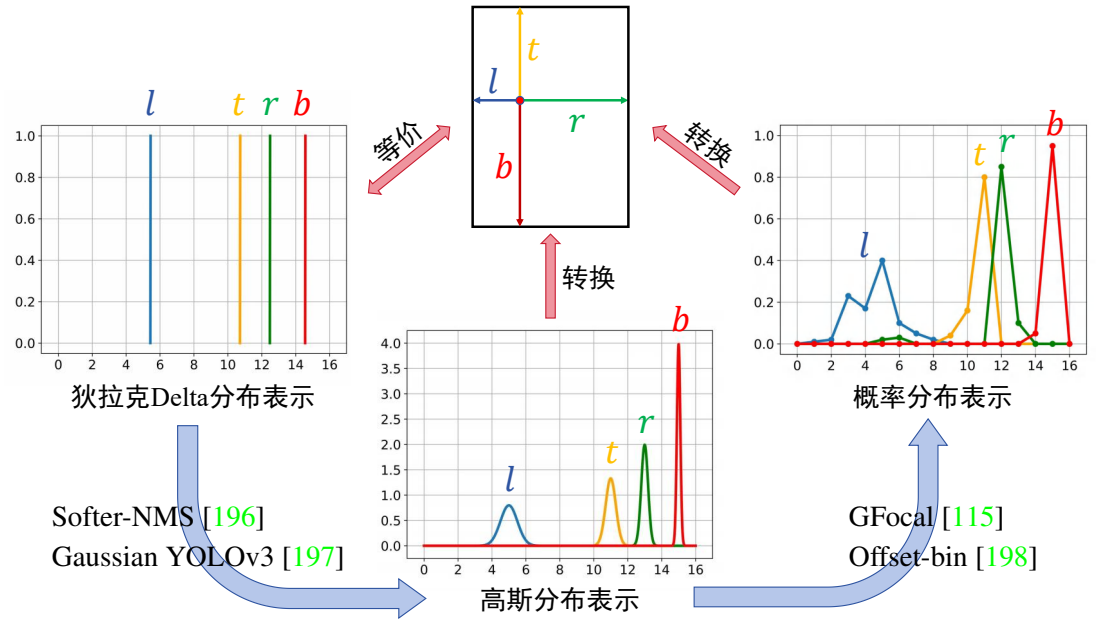


图 2.1: 三种边界框表示

进行参数更新。早期的边界框回归多使用基于 l_n 范数的损失函数来衡量预测框与真实框的误差，如 YOLOv1 [109]使用均方误差，以及 Fast R-CNN [95]中所使用的 Smooth L1损失，

$$\mathcal{L}_{\text{smooth}}(\mathcal{B}_p,\mathcal{B}_g)=\begin{cases}0.5(\mathcal{B}_p-\mathcal{B}_g)^2, & |\mathcal{B}_p-\mathcal{B}_g|<1\\ |\mathcal{B}_p-\mathcal{B}_g|-0.5, & |\mathcal{B}_p-\mathcal{B}_g|\geqslant 1\end{cases} \tag{2.1}$$

这种损失函数是直观的，易于部署。然而由于 l_n 范数损失天然地不具有尺度不变性，即便两框的重叠关系固定，损失值会随着两框的大小而发生变化，这对于越大的框而言，损失值会急剧加大。 Unitbox [199]提出最大化预测框与真实框之间的 IoU，

$$IoU(\mathcal{B}_p,\mathcal{B}_g)=\frac{|\mathcal{B}_p\cap\mathcal{B}_g|}{|\mathcal{B}_p\cup\mathcal{B}_g|}. \tag{2.2}$$

当 $IoU\to 0$ 时，预测框 $\mathcal{B}_p$ 与真实框 $\mathcal{B}_g$ 完全不重叠；而当 $IoU\to 1$ 时，代表着两框完美重叠。最初 Unitbox使用 IoU的负对数作为损失值，而后在广义 IoU [200]中，IoU损失被定义为有界形式：

$$\mathcal{L}_{IoU}=1-IoU(\mathcal{B}_p,\mathcal{B}_g). \tag{2.3}$$

显然，有界的边界框损失在调整损失权重大小时将相比于无界形式更为便利，有利于模型获得更稳健的优化效果，且尺度不变性被保持，即对 $\forall r>0$，都有

$IoU(\mathcal{B}_p,\mathcal{B}_g) = IoU(r\mathcal{B}_p, r\mathcal{B}_g)$. IoU通常还被作为检测框的度量，在评价体系中用于衡量检测框的精确度，因而最大化 IoU被认为与最大化评估指标高度相关。

然而 IoU损失的一大弊端在于当两个框不相交时，IoU恒为0，则无法给模型提供梯度回传。于是在 GIoU [200]中，研究者巧妙地添加了一个惩罚项，

$$GIoU(\mathcal{B}_p,\mathcal{B}_g) = IoU(\mathcal{B}_p,\mathcal{B}_g) - \frac{|C| - |\mathcal{B}_p \cup \mathcal{B}_g|}{|C|}, \tag{2.4}$$

再使用 $\mathcal{L}_{GIoU} = 1 - GIoU(\mathcal{B}_p,\mathcal{B}_g)$ 作为损失，其中 $\mathcal{C}$ 表示 $\mathcal{B}_p$ 与 $\mathcal{B}_g$) 的最小覆盖矩形框。 GIoU损失依然保持有界性、尺度不变性，并且对于矩形框多样化的重叠关系刻画更加细致。然而由于 GIoU的惩罚项依然建立在两个框的重叠面积的运算，只可能导致模型收敛缓慢以及回归不准确的问题。 DIoU [201]提出了最小化两个框之间的归一化中心点距离，

$$\mathcal{L}_{DIoU} = 1 - IoU(\mathcal{B}_p,\mathcal{B}_g) + \frac{d^2}{c^2}, \tag{2.5}$$

其中 $d = d(\mathcal{O}_p,\mathcal{O}_g)$ 为两框中心点距离，$\mathcal{O}_p,\mathcal{O}_g$ 分别为预测框与真实框的中心点，c 为最小覆盖矩形 $\mathcal{C}$ 的对角线长度。更进一步，研究者还总结了一个好的边界框回归损失应考虑三个重要几何因子：重叠面积、中心点距离、宽高比，于是一种完全的 IoU损失被提出，称为 CIoU损失，

$$\mathcal{L}_{CIoU} = 1 - IoU(\mathcal{B}_p,\mathcal{B}_g) + \frac{d^2}{c^2} + \frac{4}{\pi^2}(\arctan\frac{w_p}{h_p} - \arctan\frac{w_g}{h_g})^2, \tag{2.6}$$

其中 w_p, h_p 表示预测框的宽高，w_g, h_g 表示真实框的宽高。 CIoU损失综合考虑了边界框回归的三个几何因子，带来了更快的收敛与回归准确度，后续被 YOLOv4 [116]及其后续的系列所采用。

2.2.3 定位质量估计

在章节 2.2.1中，本文提到了定位模糊性广泛存在于物体的边缘，这是导致目标检测器难以学习到准确的检测框边界的原因之一。定位质量估计就是一种旨在刻画定位模糊性的方法，并期望以此改善目标检测的检测性能。早期的定位质量估计可追溯到 YOLOv1 [109]，其中研究者使用预测的物体置信度来惩罚分类得分。 2018年，IoU-Net [202]提出预测检测框与真实框的 IoU，该 IoU预测分支以预测框与真实框的实际 IoU为标签。 IoU-Net使用预测的 IoU来引导非极大值抑制算法。随后，Mask scoring R-CNN [203]提出预测物体掩码 IoU，并

同样惩罚分类得分。 FCOS [101]与 PolarMask [204]分别提出预测检测框中心度（centerness）与物体掩码中心度来惩罚分类得分，中心度的作用同样是降低定位质量较差的检测框的分类得分，以使得定位质量较高的检测结果更容易在非极大值抑制算法下被保留。

除了上述对检测框进行单值的定位质量估计，前文所提到的高斯分布边界框表示使用均值与方差来表示检测框的一条边界，其中检测框边界的方差也可视为对该边界的定位质量估计。因而 Softer-NMS [196]使用方差来作为权重，在加权平均非极大值抑制中赋予定位较准的边界更大的权重，而赋予定位较差的边界更小的权重。 Gaussian YOLOv3 [197]则沿用了过去的得分惩罚机制的思想，使用检测框边界的方差来惩罚分类得分。 GFocalv2 [129]则利用边界框的概率分布表示，提取较高的几个概率值来形成边界框的定位质量估计，再使用额外的分支将此信息注入到分类分支的学习，使得分类任务与定位任务更好地融合。

2.3 目标检测知识蒸馏

知识蒸馏（Knowledge Distillation，KD）作为一种经典的模型压缩算法，近年来已称为深度学习中的热门研究领域。知识蒸馏的基本思想是使用性能良好的大型教师网络将捕获的知识转移到小型学生网络。Logit模拟，又称分类 KD，由 Hinton等人首次提出 [205]，其中学生模型的 logit输出受到教师模型的 logit输出监督。随后 FitNet [206]指出，知识蒸馏不仅仅应该在模型的输出 logit上进行，也应该在中间隐藏层神经元上进行。 FitNet提出在师生模型的中间深层特征上施加均方误差，以对齐师生特征表示。以上两种方法，分别奠定了知识蒸馏算法的基础，其中 KD为模型输出层面的蒸馏代表，而 FitNet则成为了特征蒸馏的代表，如图 2.2所示。

知识蒸馏首次由 Chen等人 [207]引入目标检测领域，该工作以 Faster R-CNN [96]为基础，构建了一套集合了 logit模拟与特征模仿的目标检测知识蒸馏框架，并且该框架还包括一个用于定位头的伪边界框回归，用于拟合高质量的教师预测框。随后，目标检测知识蒸馏开始受到学界的关注。然而，由于 logit模仿受限于蒸馏效率较低，因而学界更多地改进主要集中于特征模仿，分为三个方面：

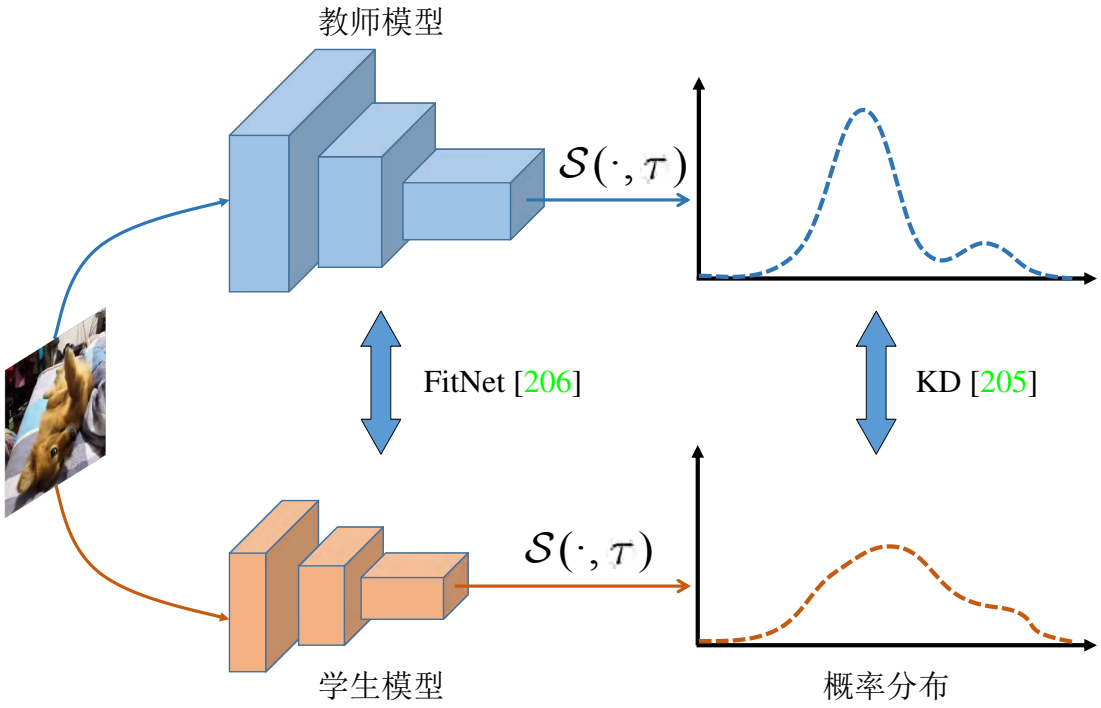


图 2.2: 知识蒸馏框架

(1) 挑选蒸馏区域：比较有代表性的方法如 Li等人 [208]提出在 Faster R-CNN的区域提案内部进行特征模仿，以期望知识传递更多地聚焦于感兴趣的区域上。FGFI [209]则选择在一些高质量的锚点位置进行蒸馏。 Guo等人 [210]注意到以往的蒸馏方法通常聚焦于靠近物体的区域而忽略了背景区域所含知识的重要性，于是研究者提出了 DeFeat蒸馏方法，其为前景和背景区域分配了不同的蒸馏损失权重，特别是能够增大背景区域的蒸馏损失，以实现更好的蒸馏效率。 Dai等人 [211]提出了通用实例选择模块，以挑选师生预测差距较大的特征点进行蒸馏。

(2) 加权蒸馏损失权重：一些研究工作则是从加权蒸馏损失的角度来调整不同区域传递知识的重要程度。较为有代表性的方法如高斯掩码加权 [212]，在预测框的中心权重大，而向外逐渐高斯衰减。 FRS [213]提出使用教师检测器最大分类得分作为蒸馏损失权重，而 PFI [214]则逐通道上使用师生检测器的分类得分 l_1 平均误差作为蒸馏损失权重。

(3) 蒸馏相关性： 与点对点的师生对之间特征模仿不同，蒸馏相关性考虑了单

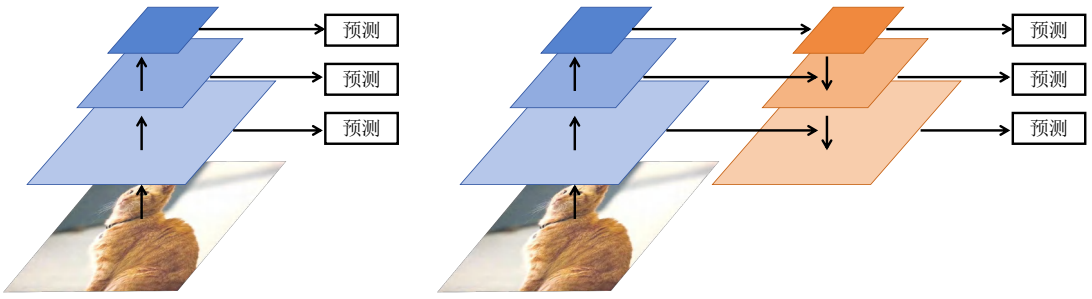
个模型自身不同特征点之间的相关性也可作为一种潜在的知识，值得被传递给学生模型。 RKD [215]展示了蒸馏特征相关性在分类任务上可取得有效的改进。GID [211]在此观点的基础上对师生预测差距较大的特征点上引入了基于相关性的知识蒸馏。 PKD [216]提出最大化师生特征的皮尔逊相关系数，该方法可等价于先对师生特征分别进行高斯标准化后再施加均方误差。

2.4 多层级学习

多层级学习，也称为颈部网络+检测头网络，是一种传统的以特征金字塔 [96, 104, 106, 110, 116, 217]的方式检测各种尺度物体的学习范式。本节将介绍目标检测领域中的一些较有代表性的颈部网络与检测头网络。

2.4.1 颈部网络

关于如何构建更强大的颈部网络，近年学界的研究取得了丰硕的进展。这其中最具代表性的成果之一当属特征金字塔 FPN （Feature Pyramid Network）[105]。 FPN 起源于一个最基本的观察，即不应当将各种大小的物体集中在一个特征层级上进行预测。在此之前，SSD [110]意识到这一基本原则，提出构建了多级预测的方式，将大物体的预测交给深层特征图，而小物体则交给浅层特征图，如图 2.3(a)所示。而后，FPN 进一步强化这一学习范式，如图 2.3(b)所示，其观察到浅层特征缺乏足够的语义信息，不利于小物体的检测，于是提出将高层特征自上而下传递至浅层，通过特征融合，来加强小物体的检测。 此后这类



(a) SSD 预测范式　　(b) FPN 预测范式

图 2.3: 目标检测的预测范式

特征金字塔网络被学界统一称为 neck （颈部）网络，它们是主干网络与预测头之间的桥梁，起到特征加工与分流的作用。较为著名的 neck网络设计还包括使用尺度变换的 STDN [218]，每四张小特征图拼图成一张大特征图，从而使其

可以支持更强大的主干网络 DenseNet [219]。 PANet [220]在 FPN的基础上进一步增加了自下而上的路径，以达到更好融合特征的效果。 DSSD [221]利用反卷积 [222]构建了具有丰富语义信息的特征金子塔。 NAS-FPN [223]则尝试构建一种神经网络搜索方案的 neck网络，搜索出的 neck网络结构交错复杂，但因其高度自主学习，对不同场景适应性较强。 ASFF [224]采用了加权聚合的原理，引入了可参与训练的加权因子来体现不同层级特征图的重要性。 Tao Kong [225]等人提出了一种非线性聚合策略，其观察到以往的 FPN 本质上是一种线性聚合策略，设特征金字塔的不同层特征图为 $X_1, X_2, \cdots, X_n$，则 FPN 的聚合可以表达为如下的过程：

$$\begin{aligned} X'_n &= X_n \\ X'_{n-1} &= a_{n-1}X_{n-1} + b_{n-1}X'_n \\ X'_{n-2} &= a_{n-2}X_{n-2} + b_{n-2}X'_{n-1} \\ &\vdots \\ X'_1 &= a_1X_1 + b_1X'_2 \end{aligned} \tag{2.7}$$

其中 $a_i, b_i, i = 1, 2, \cdots, n$ 是特定的线性操作，如不带有激活函数的卷积，双线性插值上采样，反卷积等。那么在 [225]一文中，作者借鉴了 SENet [226]的自注意力思想，来自适应地聚合特征，分配至不同的层级。 EfficientDet [227]中的 Bi-FPN则受到 EfficientNet [228]的复合缩放原则启发，分别从通道、分别率、堆叠块数三个层面进行复合缩放，可获得较有性价比的效能平衡。众所周知，当网络块堆叠得越多，通常越有利于精度提升。 Neck网络的另一种有趣改进是无限堆叠，i-FPN [229]，其主要受到深度均衡模型 DEQ [230]的启发，尝试以较小的代价来模拟网络块无限堆叠时所达到的特征均衡态的性能。 QueryDet [231]添加了 P2层级、查询头和稀疏卷积以权衡速度和精度，但它需要一些额外的手工设计来使用稀疏卷积操作，并且必须随着预测图的变化再次为损失函数寻找更好的超参数。 YOLOF [232]构建了一个单层级密集目标检测器。虽然 YOLOF成功减少了计算负担，但它依赖于一些针对单层级模型的定制设计，例如堆叠膨胀卷积块和均匀标签匹配。这使得将单层级模型难以推广到流行的多层级模型上。

2.4.2 检测头网络

检测头网络通常用于进一步细化上游网络的特征。其典型组件是堆叠

的卷积操作。卷积的数量通常从1个（RPN [96]）到6个（TOOD [131]）不等，其中4个卷积层被广泛采用，例如 RetinaNet [106]、FCOS [101]等。一些目标检测器在检测头网络中采用全连接层（FC layer），例如知名的 R-CNN系列检测器 [96, 102, 153, 233]。另一个特殊检测器是 Double-Head [234]，它根据经验观察到全连接层适合分类任务，而定位任务更青睐于卷积层。 Dynamic Head [235]在检测头网络中考虑了3种注意力机制，即尺度、空间和任务感知。 GFocal [115]提出联合优化分类和定位，并删除了 FCOS提出的中心度分支。 DDQ-FCN [236]在检测头网络中集成了通道融合，而 GFocalV2 [129]添加了 FC模块来预测定位质量估计。 Softer-NMS [196]利用分数投票 NMS（非最大抑制）来对检测框进行精细化处理，而 VFNet [130]提出了使用星形框特征表示来加强定位能力。一些方法则提出了更好的边界框表示来捕捉定位模糊性，例如高斯分布表示 [196, 197]和概率分布表示 [115, 198]，从而提高了定位质量。还有一些方法可以在不损失推理效率的情况下提高检测性能，例如标签分配算法（FreeAnchor [237]、ATSS [114]、PAA [137]、OTA [138]、DW [132]、SELA [238]）、损失函数（GHM [154]和基于 IoU的损失函数 [199–201, 239, 240]）、以及知识蒸馏（LD [241, 242], FGD [243], PKD [216], CrossKD [244]）。

在上述方法中，检测头网络中的操作通常是跨特征层级并行执行的。以前的工作很少关注不同特征层级之间的效率不平衡问题。在本文中，将重新思考了多层级学习，并提出了一种新的 SlimHead检测头网络来平衡浅层级和深层级之间的计算复杂性。

2.5 数据集与评估

目标检测器的应用离不开数据集与评估两方面的支持，前者为目标检测器提供应用场景的泛化迁移能力，而后者为目标检测器提供性能度量指标，为更好地高效能优化目标检测器而服务。

2.5.1 目标检测数据集

目标检测所使用的数据集包括学界较为知名的基准数据集，如图 2.4所示，也包括一些为特定场景特定任务所开发的数据集，如图 2.5所示。下面将详细予以介绍。

PASCAL VOC [36] MS COCO [37] DOTA [29] CrowdHuman [245]

图 2.4: 学界通用目标检测基准数据集示例图

PASCAL VOC

PASCAL VOC [36]是目标检测界最知名、使用最广泛的数据集之一，多年被作为一项世界级的计算机视觉挑战赛。目标检测是其主流赛道之一，涵盖20个生活自然场景下的类别，如人、车、猫、狗、飞机等等。 PASCAL VOC含两个最著名的子集，VOC 2007与 VOC 2012。目前学界最常用的训练与测试基准为 VOC 07+12协议。该协议使用 VOC 2007 trainval集与 VOC 2012 trainval集作为训练集，共计16551张图像，使用 VOC 2007 test集作为评估集，共计4952张图像。

MS COCO

随着神经网络逐渐加深与加宽，模型的性能愈发强大，数据集规模较小的PASCAL VOC已日渐难以满足学界的研究需求。建立一种更大型、更困难、更具有挑战性的检测数据集已成为目标检测发展的必然要求。 MS COCO [37]正是为此应运而生，其相较于 PASCAL VOC拥有更大规模的图像数量，更多的自然场景下的类别，共80个类，以及包含更多的物体遮挡场景与小目标。 MS COCO的主流训练与测试基准采用 COCO train2017为训练集，共计11万8千余张图像，使用 COCO val2017作为评估集，共计5000张图像。此外，MS COCO官方团队还提供了未公开标签的测试集 COCO test2017，共两万余张图像，为学界提供了一个比赛级的对比环境。研究者可将 COCO test2017的推理结果打包上传

至 COCO官方评估服务器中进行评估，得到评估指标。

DOTA

DOTA [29]是遥感影像中最经典的数据集之一，其常作为有向目标检测的评估数据集。 DOTA含三个常用版本：DOTA-v1.0、DOTA-v1.5以及 DOTA-v2.0。DOTA-v1.0包含18万余个遥感物体，共计2806张大尺度图像中，覆盖15个类别，含交通工具、游泳池、飞机、船只等等。 DOTA-v1.0的训练、验证、测试构成为3:1:2。本文将主要使用 DOTA-v1.0版本进行实验与分析，以下简称为 DOTA。在使用 DOTA时，需要事先将大尺度图像裁剪为 600×600 大小的切片，相邻切片重叠像素为150，在放大至 800×800 大小。

口罩　　水果　　安全帽

图 2.5: 特定任务场景的目标检测数据集示例图

CrowdHuman

行人检测是目标检测的一个重要应用分支，也是由来已久的经典计算机视觉任务之一。 CrowdHuman [245]是一个涵盖各种场景的大规模行人检测数据集。其每张图片中平均出现22.6个行人目标，这意味着 CrowdHuman中的行人场景更加拥挤，互相遮挡的案例也更丰富。本文将主要使用 CrowdHuman train集

（15K张图片）来进行训练，评估集为 CrowdHuman val集（4.4K 张图片）。

口罩数据集

随着新冠肺炎疫情在全球肆虐，口罩检测是一种广泛且必要的视觉应用，常部署于机场、医院、大型超市等人流量大的公共场所。口罩检测数据集 [246]包含5865张用于训练的图像和1035张用于评估的图像。该数据集含有两个类别，人脸与口罩，以识别图像中所存在的人脸，以及其是否佩戴了口罩。

水果数据集

水果检测是一种广泛应用于工业流水线分拣、零售商品分类的计算机视觉应用。水果检测数据集 [247]包含3836张图像用于训练，以及639张图像用于评估模型性能。该数据集含有11种常见水果类别，例如苹果、葡萄、柠檬、西瓜等。

安全帽数据集

安全帽检测是一款安全视觉应用程序，常用于建筑工地检测工人和访客是否戴着安全帽。安全帽数据集 [248]包含15887张用于训练的图像和6902张用于评估的图像。与口罩数据集类似，安全帽数据集含有两个类别，人头与安全帽，以识别图像中所存在的人头，以及其是否佩戴了安全帽。

2.5.2 目标检测评估指标

给定一个目标检测器，当前学界对其主要有三个维度的评估：精确度、召回率、速度。一个高效能目标检测器，意味着其拥有高精确度、高召回率与高速度。而在目前主流的评估体系中，MS COCO [37]的平均精确度（Average Precision，AP）是最具有代表性的指标之一，评测过程具体如下：

- **得分排序与过滤**：对检测框的分类得分（或置信度）降序排列，使用阈值 r，例如0.01，过滤掉低分类得分框。
- **判断检测框的属性**：对于每一个 IoU阈值 $\mu \in \{0.5, 0.55, 0.6, \cdots, 0.95\}$，依据混淆矩阵逐个判断检测框的属性是四个类型中的哪一种，如表2.1所示。判断依据为检测框与真实框之间的 IoU是否大于给定阈值 μ。当有多个检测框与某一个真实框的 IoU都超过阈值 μ 时，则会选择最大分类得分框为TP。（注：在上述筛选过程中，所有被淘汰的检测框，无论是低分类得分

框，还是 IoU小于阈值 μ 的检测框，亦或者重复对应同一真实框的检测框都会被标记为 FP。）

表 2.1: 混淆矩阵。

混淆矩阵		真实值	
		阳性	阴性
预测值	正确	TP （真阳性）	FP （假阳性）
	错误	FN （假阴性）	TN （真阴性）

- **绘制 PR曲线**：对每一个类别，从最大置信度检测框开始，逐个添加检测框，计算精确度 $P=TP/(TP+FP)$ 与召回率 $R=TP/(TP+FN)$，绘制出PR曲线 $P(R)$，经过去尖峰化后，得到11个评估点 $R\in\{0,0.1,0.2,\cdots,1\}$，如图2.6所示。

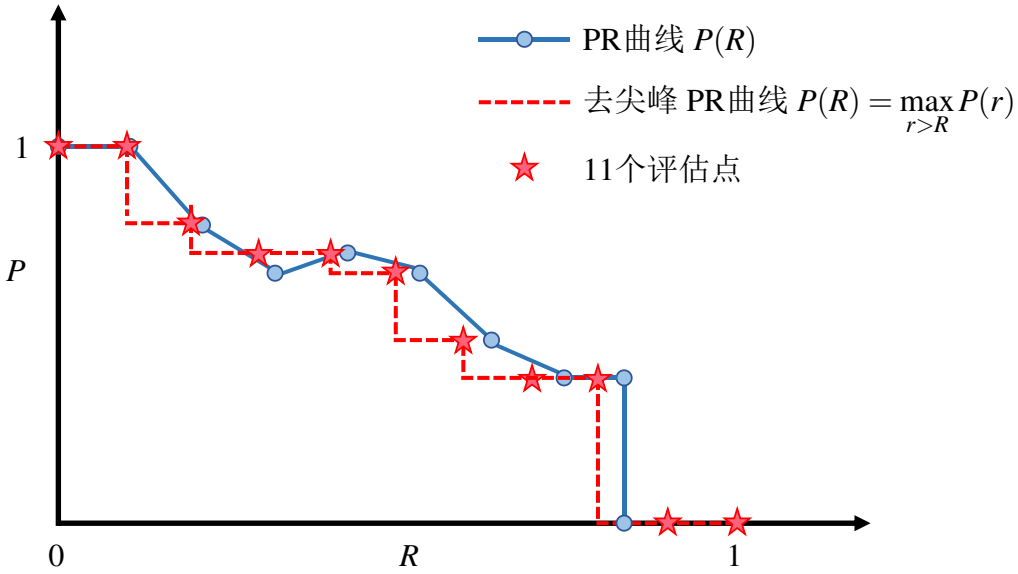


图 2.6: PR曲线与 COCO 11点评估法

- **计算 mAP**：将以上11个评估点的 P值求和平均，便得到单类别的平均精确度 AP_c，再对所有类别取平均得到 mAP@μ。
- **COCO评估指标结果**：十个 IoU阈值下的 mAP平均值：

$$AP=\frac{1}{10}\sum_{\mu}\text{mAP@}\mu,\quad \mu\{0.5,0.55,\cdots,0.95\}. \tag{2.8}$$

另有5个指标为 AP_{50}（mAP@0.5），严格定位指标 AP_{75}（mAP@0.75），小物体平均精确度 AP_S，中物体平均精确度 AP_M，大物体平均精确度 AP_L，其中大中小物体分别定义为真实框面积 $S > 96^2$，$32^2 < S \leq 96^2$, $S \leq 32^2$。

除了以上能够综合反映目标检测器精确度与召回率的平均精度（AP）以外，对于检测速度的评估指标，学界常采用 FPS （Frames Per Second），每秒帧数，作为衡量单位，以评估检测器的运行效率。在计算检测速度时，一般将2000张图像逐一送进检测器进行推理，统计推理过程平均耗时。其中单张图像的推理耗时则统计了该图像被送进检测器模型直至推理得出检测框的所用时间。一般而言，学界以 FPS等于30作为实时性目标检测的分界点。

第三章 区域评估：揭示面向空间均衡的新阻碍

目标检测器的一个根本局限性是，它们会遭受“空间偏差”的影响，尤其是在检测图像边界附近的目标时，性能会显著下降。长期以来，目标检测领域一直缺乏有效的方法来测量和识别空间偏差，对于这种偏差的来源和程度也知之甚少。为了解决这个问题，本章通过从传统评估扩展到更通用的评估，提出了一种新的目标检测评估方法，称为区域评估。该方法通过测量不同区域的检测性能，产生一系列区域精度 (Zone Precisions, ZPs)。本章首次提供了数值结果，结果表明，目标检测器在不同区域的性能表现非常不均衡。更令人惊讶的是，检测器在图像96%的边界区域的性能甚至达不到 AP值（平均精度, 通常被认为是整个图像区域的平均检测性能）。为了更好地理解空间偏差，本章还将进行一系列启发式实验。实验结果有力排除了关于空间偏差的两个直观猜想，即目标尺度和目标的绝对位置几乎不会影响空间偏差。本文发现，关键在于不同区域目标数据模式之间，人类难以察觉的差异，这些差异最终导致了区域之间明显的性能差距。基于这些发现，本章最终讨论了目标检测的未来方向，即空间不均衡问题，旨在追求在整个图像区域内实现均衡的检测能力。最后，通过广泛评估10种流行的目标检测器和5个检测数据集，本章揭示了目标检测器的空间偏差问题。

3.1 引言

目标检测在过去二十年中取得了令人瞩目的进展 [96, 104, 110, 117]。虽然目标检测器的优化流程已被充分探索，但它们在局部图像区域内的行为仍然是一个谜。目标检测器的空间鲁棒性尤其重要 [167]，因为目标可能出现在任何位置，并且所有目标都应该被良好地检测到。这对于安全视觉安全应用尤为重要，例如，火灾/烟雾检测 [249, 250]、自动驾驶中的汽车防撞 [197, 251]、人群计数和定位 [252–255]、智能监控系统中的武器检测 [256, 257]、以及商店盗窃检测 [258]等，在这些应用中，边界区域占据了图像区域相当大一部分。

不幸的是，检测器实际上无法在空间区域内均匀地执行检测，通常在图像边界附近表现出明显的性能下降。这种现象，本文称之为“空间偏差”，是目

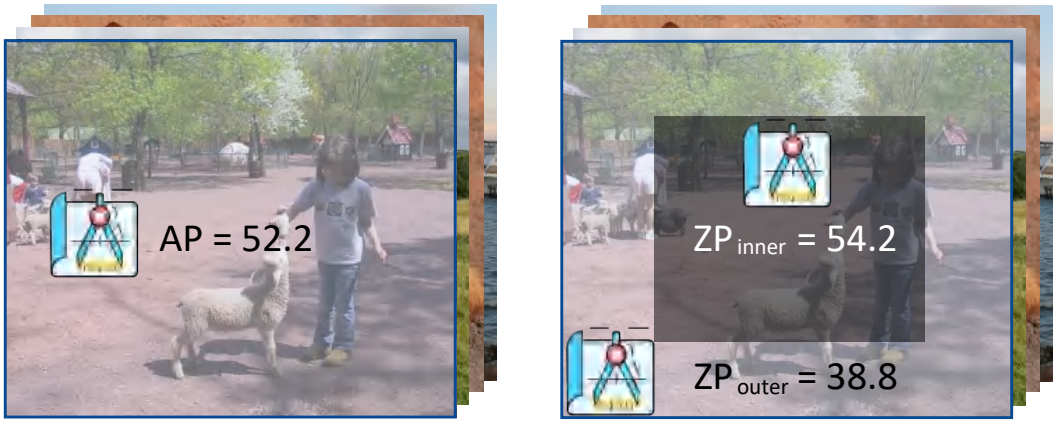


图 3.1: 传统的评估方法衡量整个图像区域的检测性能，但它忽略了对局部区域的测量，难以反映空间偏差。本文的区域评估 (ZP，区域精度，即区域内约束的平均精度) 弥补了这些问题，表明区域之间存在很大的性能差距。结果由 GFocal [115]在 VOC 2007测试集 [36]上报告。

标检测中一个天然的障碍，但在很长一段时间内都被检测社区所忽视。忽视这个问题可能会导致严重的安全隐患和重大财产损失的风险。例如，火灾探测器可能擅长检测中心区域的火灾，但会失去检测图像边界区域火灾的能力。这样的火灾报警系统是不可靠的，因为镜头中心区域仅占图像区域的一小部分。卷积神经网络 (CNN) 在空间鲁棒性方面的一些最新突破 [156, 157, 159, 259]正朝着难以捉摸的平移不变性方向发展，其基础是理解小的图像变换（例如，颜色抖动、平移）如何影响分类模型的分类精度。有研究发现，即使对于同一个目标，分类器也会随着目标空间位置的变化而做出完全不同的预测结果 [159]。而在图像分类之外，本章深入研究了目标检测中的空间偏差，揭示了现代目标检测器的局限性。

多年来，一直存在一个开放性问题，即缺乏一种有效的方法来测量和识别空间偏差。传统的评估方法，即 AP指标，衡量的是整个图像区域的检测性能，这并没有为检测器的空间鲁棒性提供任何指导，人们也很难知道检测性能在一个局部区域内的情况，以及下降了多少。因此，评估协议显得尤为重要，因为它可以提供机会更好地理解空间偏差，并为进一步构建方法论提供工具。为此，本文提出了一种系统的方法，称为区域评估，来分析现代目标检测器中是否存在空间偏差，以及如果存在，偏差有多大。具体而言，本文将传统的全图评估扩展到更通用的评估。本文计算指定区域内的通用平均精度（AP） [37]，从而

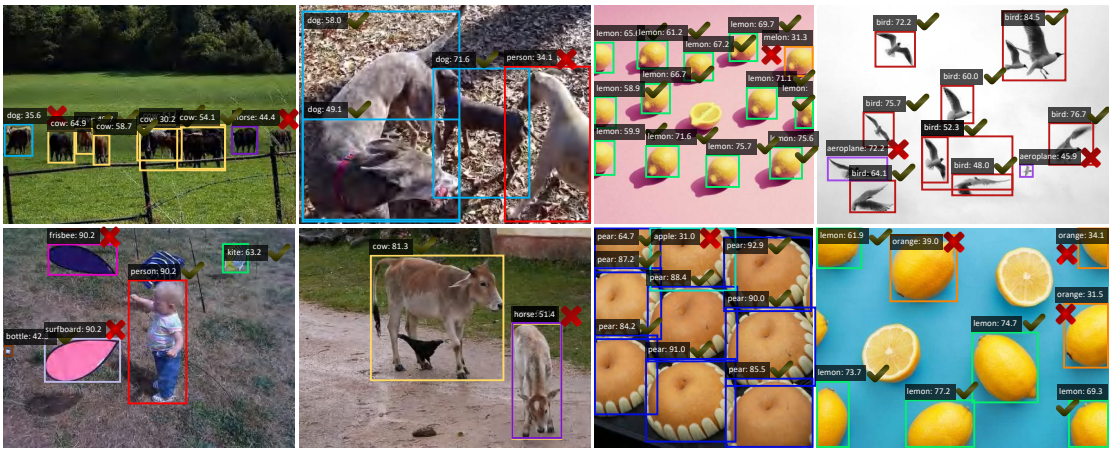


图 3.2: 检测器在检测边界目标时不太理想。可视化结果由 GFocal [115]报告。放大以获得更好的观看效果。

得出区域精度 (Zone Precision, ZP)。在评估期间，在测量某一个区域内的检测性能时，将仅考虑中心点位于该区域内的框。借助这些辅助指标，本章首次提供了数值结果，明确揭示了目标检测器实际上在不同区域之间存在相当大的空间偏差。如 图 3.1所示，内部区域和外部区域之间的 ZP 差距为 15.4。表面上看，这种性能差距可以推断出，检测能力似乎与目标的绝对位置高度相关。然而，当我们移动图像中的目标时，这种看似合理的推测在实践中存在许多根本上的不一致之处。目前学界尚没有任何对边界区域性能下降的令人满意的解释（见 图 3.2）。因此，本章希望阐明目标检测器中空间偏差的存在和主要来源。最后，本章提出了未来目标检测研究的一个重点：走向空间均衡。

本章的贡献主要包括一个全新的区域评估方法、对空间偏差的三个探索性实验、一个全新的研究问题的建立与解决方法的提出，以及对现代目标检测器的全面评估。

- **一个新的评估方法：区域评估。** 本章节首先提出了一个全新的区域评估方法来从局部视角对检测器进行评估。通过区域划分，检测器将在数个区域上进行性能评估，得到一系列区域指标以及区域指标的方差。这对于度量空间偏差现象以及相应解决方法的探索有着重要指导意义。

- **空间偏差的三个可能因素：** 1. 目标尺度是否在中心区域性能中起关键作用？ 答案是否定的。虽然大型目标相对频繁地出现在中心区域，但本文观察到，当我们很大程度上消除目标尺度的影响时，不同区域之间的区域性

能仍然非常不均匀。空间偏差和目标尺度之间仍然难以建立必然的联系。

2. 检测器是否根据目标的绝对空间位置产生中心区域性能？ 答案是否定的。检测性能几乎与目标的空间位置无关。本章观察到，当目标移动到中心区域时，这在统计学上不会导致检测质量的提高。

3. 什么才是真正决定目标检测器空间偏差的因素？ 本章的分析揭示了强有力的证据，表明空间偏差与区域之间目标数据模式的差异高度相关。具体而言，中心目标和边界目标之间在目标数据模式上存在差异。因此，如果一个目标是从中心区域风格的数据分布中采样，而不是从边界区域风格的数据分布中采样，则可以更好地检测到该目标。

- **一个新的未来方向：空间失衡问题与空间均衡学习。** 本章更进一步，提出了一个迫切需要解决的实际问题，即空间不均衡问题。在这种问题设置下，目标检测器将空间均衡作为重要目标之一，这对鲁棒检测具有至关重要的意义。面对这一挑战，本章节还提供了首次尝试，即空间均衡学习，以实现空间均衡目标检测。

- **全面的评估。** 本章提供了几种代表性目标检测器的广泛评估和比较。本章通过实验揭示：1) 空间偏差在各种目标检测器和数据集中非常普遍。2) 检测器的空间均衡性差异很大。特别是，本章节将展示稀疏检测器在中心区域表现更好，而单阶段密集检测器在边界区域表现更好。3) 提出的空间均衡学习能够缓解空间不均衡问题。

3.2 区域评估

本节将传统的目标检测评估扩展到更通用的区域评估。给定一个测试图像 I 和一组评估指标 $\mathcal{M}$，经典的评估方法同时计算整个图像中所有检测结果和真实标签的指标。 $\mathcal{M}$ 中的元素可以是 COCO AP（平均精度） [37]，10个 IoU阈值上的 AP，或小/中/大目标的 AP，这些都在目标检测界被广泛采用。这些传统指标衡量了整个图像区域的检测性能，但没有考虑目标检测器的空间鲁棒性。

区域指标

令 $S=\{z^1,z^2,\cdots,z^n\}$ 是一个区域划分，使得 $I=\bigcup_S z^i$ 且 $z^i\cap z^j=\emptyset$，$\forall z^i,z^j\in S$，$z^i\neq z^j$. 本文通过仅考虑中心位于区域 z_i 内的真实目标和检测结果来衡量特定区

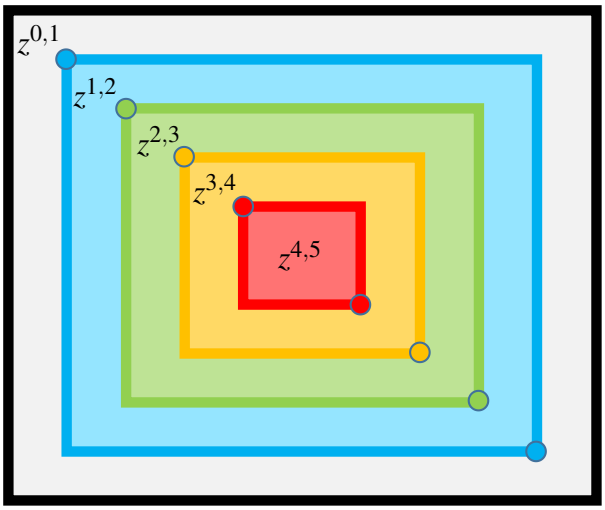


图 3.3: $n=5$ 时评估区域的定义

域 z_i 的检测性能。然后，对于任意评估指标 $m \in \mathcal{M}$，评估过程与传统方式保持一致，产生 n 个区域指标，每个指标记为 m^i. 我们称 $m^S = \{m^1, m^2, \cdots, m^n\}$ 为区域划分 S 的区域指标序列。

环形区域

在实践中，中心化的摄影师偏差在视觉数据集中普遍存在 [36,37,260–263]。如果人们想要一个具有全面检测能力的检测器，则评估区域可以设计成一系列环形区域：

$$z^{i,j} = R_i \setminus R_j, \quad i < j, \tag{3.1}$$

其中 R_i 表示一个中心区域，由下式给出：

$$R_i = \text{Rectangle}(p,q) = \text{Rectangle}((r_i W, r_i H), ((1-r_i)W, (1-r_i)H)), \tag{3.2}$$

其中 $\text{Rectangle}(p,q)$ 表示左上角坐标为 p，右下角坐标为 q 的矩形区域。W 和 H 表示图像的宽度和高度，$r_i = \frac{i}{2n}, i \in \{0,1,\cdots,n\}$ 控制矩形的大小。评估区域如图 3.3所示，这里 $n=5$。本文将区域 $z^{i,j}$ 中的平均精度（AP）表示为 $\text{ZP}^{i,j}$。通过这种方式，传统评估成为了本文区域评估中的一个特例，因为可以很容易地得到 $\text{AP} = \text{ZP}^{0,n}$。

其他区域划分

传统评估可以灵活地为不同的应用场景选择不同的 IoU阈值。对于那些需

要精确定位框的应用，可以选择严格的 IoU阈值，例如，$IoU = 0.75$ (AP_{75})。对于那些对框定位要求较低的应用，AP_{50} 就足够了。例如，旋转目标检测方法通常报告 AP_{50} [79, 80, 264]。类似于 AP指标，用户可以根据自己的应用灵活地设计各种区域划分。如果用户关心全方位综合检测能力，环形区域划分将是一个不错的选择。如果用户关心某些感兴趣的区域，则可以自定义评估区域。本章的实验部分将展示另外两种特殊区域划分的评估结果。一种是条形区域，即沿 x 轴的5个区域和沿 y 轴的5个区域，另一种是 11×11 的正方形区域。重要的是，由于区域划分保持一致，因此目标检测器之间的比较仍然是公平的。此属性有助于我们观察不同检测器在感兴趣区域的性能，以便可以根据实际应用需求选择检测器。在后续章节，本文将展示稀疏检测器在中心区域表现更好，而单阶段密集检测器在边界区域表现更好。

衡量区域指标的离散幅度

由于检测性能在不同区域之间变化，本文进一步引入一个额外的指标来衡量区域指标之间的离散幅度。给定特定区域划分 S 的所有区域指标 m^S，本文计算区域指标的方差 $\sigma(m^S)$。理想情况下，如果 $\sigma(m^S) = 0$，则目标检测器**在当前区域划分下**的泛化能力达到完美的空间均衡。在这种情况下，目标可以被很好地检测到，而不会受到其数据模式的影响。还值得一提的是，空间偏差是目标检测器的一种外部表现，ZP方差只能反映给定区域划分的空间均衡性。换句话说，有以下三个概念。

1. 令 S 为区域划分，如果 $\sigma(m^S)$ 充分小，则检测器的空间均衡性对于 S 来说是良好的。
2. 令 S_1, S_2 为两个具有 n 个区域的不同区域划分。如果 $\sigma(m^{S_1}) < \sigma(m^{S_2})$，则认为检测器在区域划分 S_1 的方式上更具有空间均衡性。
3. 检测器没有空间偏差，表明对任意一个 S，$\sigma(m^S)$ 都充分小。

3.3 揭示空间偏差的存在性

本节将对10个流行的目标检测器和5个目标检测数据集进行全面的评估。

3.3.1 实验设置

检测器和指标

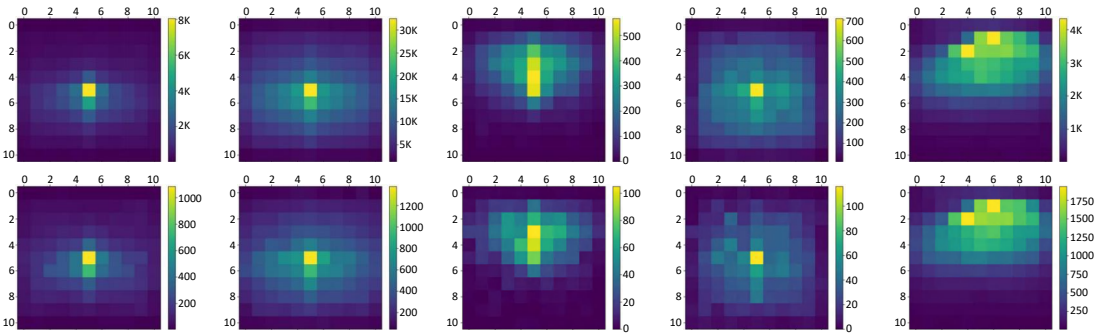


图 3.4: 5个目标检测数据集中的摄影师偏差。该图统计了所有真实框的中心点。图像被划分为 11×11 个区域。第一行：训练集。第二行：评估集。

本节评估的所有目标检测器都可以从 MMDetection [265]或其官方网站下载。本节遵循标准的 MS COCO [37]平均精度评估协议。为了全面评估检测器，本节报告了各种指标，包括5个 ZP、ZP的方差以及传统指标 AP。

数据集

本节使用 PASCAL VOC [36]、MS COCO [37]、口罩检测数据集、水果检测数据集、以及安全帽检测数据集进行评估。所有数据集皆为公开可用的，均可从其官方网站或 Kaggle下载。 5个数据集的目标分布于图 3.4中展示。关于所使用的数据集的具体描述，请参阅章节 2.5。

空间均衡学习的设置

对于空间均衡学习评估，该实现基于 MMDetection [265]框架，并且消融研究在经典的单阶段密集回归检测器 GFocal [115]上进行，使用 ResNet [108]主干网络和 FPN [105]颈部网络。本文对 PASCAL VOC和3个应用数据集使用 ResNet-18，对 MS COCO采用 ResNet-50。根据 GPU的数量，学习率通过线性缩放规则 [266]进行线性缩放。所有实验的训练 epoch都设置为12。本文在公式 3.6中设置 $\gamma = 0.2$，并且为了公平比较，所有其他超参数均保持不变。

3.3.2 对各种目标检测器进行区域评估

尽管传统评估为检测器的整体性能提供了良好的指导，但对于检测器的空间偏差以及位置和程度知之甚少。在这里，本文选择了各种目标检测器，它们具有不同的检测流程，但具有相同水平的传统指标。它们是流行的、具有代表性的，并且被认为是现代目标检测的基石：单阶段密集检测

器（RetinaNet [106]、GFocal [115]、VFNet [130]、YOLOv5 [217]）、多阶段由密集到稀疏的检测器（R-CNN系列 [96, 102, 233]）和稀疏检测器（DETR系列 [117, 118]与 Sparse R-CNN [121]）。定量结果在表 3.1中报告。

表 3.1: **对现有流行目标检测器进行区域评估：** 报告了5个 ZP、ZP的方差和传统指标 AP。结果在 COCO val2017上报告。 **R**: ResNet [108]. **X**: ResNeXt-32x4d [267]. **PVT-s**: Pyramid vision transformer-small [268]. **CNeXt-T**: ConvNext-T [269].

检测器	AP	方差	$ZP^{0,1}$	$ZP^{1,2}$	$ZP^{2,3}$	$ZP^{3,4}$	$ZP^{4,5}$
DETR (**R**-50) [117]	40.1	26.9	29.8	36.2	39.8	39.1	45.7
RetinaNet (**PVT-s**) [268]	40.4	19.7	30.8	36.9	39.0	37.4	44.6
Cascade R-CNN (**R**-50) [102]	40.3	18.7	30.9	36.6	39.2	38.6	44.2
GFocal (**R**-50) [115]	40.1	16.9	31.1	37.5	39.4	38.5	43.8
Cascade Mask R-CNN (**R**-101) [102]	45.4	22.4	34.7	41.6	44.3	44.4	49.1
Sparse R-CNN (**R**-50) [121]	45.0	21.6	35.8	41.9	43.4	44.0	50.3
YOLOv5-m [217]	45.2	12.9	36.0	42.3	44.5	43.2	46.7
Deform. DETR (**R**-50) [118]	46.1	23.2	36.3	42.6	45.6	45.1	51.2
Sparse R-CNN (**R**-101) [121]	46.2	21.2	36.9	42.9	44.9	44.7	51.3
Cascade Mask R-CNN (**X**-101) [102]	46.1	21.1	36.1	42.0	44.8	45.9	49.9
Mask R-CNN (**CNeXt-T**) [269]	46.2	17.6	36.7	41.9	44.5	43.6	49.7
GFocal (**X**-101) [115]	46.1	15.7	37.0	43.5	45.0	44.4	49.3
VFNet (**R**-101) [130]	46.2	15.6	36.7	43.0	45.0	44.5	48.8

有以下几个有趣的观察结果：

1. *空间偏差非常普遍。*由表 3.1可以看出，所有检测器都显示出明显的中心化区域性能，即在中心区域 ($z^{3,4}, z^{4,5}$) 表现良好，但在边界区域 ($z^{0,1}, z^{1,2}$) 表现不佳。这证实了空间偏差的存在和普遍性，并且本文首次成功地量化了图 3.2中所示的目标检测器的缺陷。
2. *目标检测器的空间均衡性差异很大。*ZP方差存在12.9到26.9的巨大差距。特别是稀疏检测器，例如 DETR系列和 Sparse R-CNN倾向于产生较大的ZP方差，而单阶段密集目标检测器在空间均衡性方面表现更好（较低的ZP方差）。这表明基于稀疏检测器在空间均衡性方面与基于密集回归的检测器不一致。本文推测这可能归因于自注意力机制捕获的全局信息。稀疏检测器首先通过 CNN提取特征，然后通过一系列注意力模块处理特征。由于训练更具动态性，中心目标可能会受到更多关注。显然，可以得出结

论，必然存在一些因素导致了检测器之间不同的空间均衡性，包括但不限于神经网络架构设计、优化和训练策略。然而，目前尚不清楚哪些组件或算法设计对空间偏差有影响。本文相信，未来对该主题的进一步研究将很有趣，并且该研究很有可能找到解决空间失衡问题的关键。

3. *传统评估未能捕捉到空间偏差*。注意到 GFocal（R-50）和 DETR（R-50）实现了相同的 AP分数40.1。然而，传统指标并没有提供关于某个区域的任何检测性能的信息。本文的区域评估表明，GFocal在边界区域 $z^{0,1}$ 和 $z^{1,2}$ 中表现更好，而 DETR则在区域 $z^{2,3}$、$z^{3,4}$ 和 $z^{4,5}$ 中表现更好。类似地，Deformable DETR（**R**-50） [118]实现了与 GFocal（**X**-101）相同的 AP。区域评估表明，Deformable DETR在中心区域 $z^{3,4}$、$z^{4,5}$ 的表现明显优于 GFocal，而在边界区域 $z^{0,1}$ 与 $z^{1,2}$ 中表现较差。而这些性能上的差异被传统评估所掩盖，人们无从得知目标检测器在不同区域上的性能优劣性。此外，有趣的是，AP指标正好介于 $ZP^{3,4}$ 和 $ZP^{4,5}$ 之间，这表明在96%的图像区域中，目标检测器的检测性能实际上低于 AP值。

启示：以上结果揭示了目标检测器的性能特性，这有助于人们更好地理解目标检测器的行为，并鼓励在实际应用场景部署时重新考虑检测器的选择。此外，对于检测器中的哪些组件导致了这些性能差异同样值得未来研究。

3.3.3 对各种数据集进行区域评估

3.2报告了 PASCAL VOC 07+12、MS COCO val2017和3个应用数据集的定量检测结果。有以下几个观察结果：

1. 可以看出，检测性能在不同区域之间变化。最靠近图像边界的区域，即 $z^{0,1}$，始终具有最低的检测性能。相比之下，中心区域 $z^{4,5}$ 在几乎所有这些情况下都具有最高的性能。
2. 一个代表性的例子，安全帽数据集，其 ZP方差仅为3.0。这表明安全帽数据集在环形区域划分的情况下实现了最佳的空间均衡性，而其他数据集则存在明显的空间不均衡问题。例如，在 PASCAL VOC上的 ZP方差为53.6，在水果数据集上为56.2。
3. 如果切换到其他区域划分，例如，沿 x 轴的5个条形区域和沿 y 轴的5个条形区域（见图 3.5(a)(b)），它们的空间均衡性会发生变化。在表3.3 中，在沿 y 轴的5个区域的情况下，口罩和安全帽数据集的 ZP方差分别增加

表 3.2: **对各种数据集进行区域评估：** VOC 07+12、COCO 2017和3个应用数据集（口罩检测、水果检测和安全帽检测）的区域评估。该表报告了5个 ZP、ZP方差和传统指标 AP。结果在 GFocal [115]上报告。

数据集	AP	方差	$ZP^{0,1}$	$ZP^{1,2}$	$ZP^{2,3}$	$ZP^{3,4}$	$ZP^{4,5}$
VOC 07+12	52.2	53.6	34.3	39.6	42.5	46.6	56.1
COCO 2017	40.1	16.9	31.1	37.5	39.4	38.5	43.8
人脸口罩	71.3	13.1	60.4	67.1	69.0	68.8	70.9
水果	76.6	56.2	60.8	69.9	71.2	75.3	83.8
安全帽	49.7	3.0	45.9	47.9	50.3	50.6	47.8

VOC	38.2	46.7	55.5	45.1	42.9
COCO	35.0	39.1	42.5	39.3	35.9
口罩	63.4	68.3	72.9	71.0	66.4
水果	72.2	72.8	80.4	76.2	69.7
安全帽	54.0	51.5	47.8	48.7	47.3

(a) 沿 x 轴的5个区域

VOC	COCO	口罩	水果	安全帽
36.1	30.6	69.1	64.5	44.7
36.9	34.8	69.7	70.7	52.8
52.5	39.4	69.8	77.8	49.8
47.4	41.4	63.8	74.3	44.4
36.2	36.3	53.4	68.8	36.6

(b) 沿 y 轴的5个区域

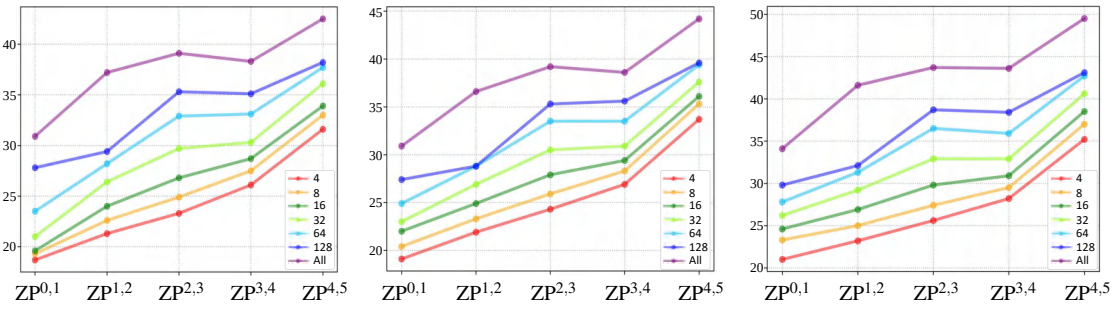
图 3.5: 区域划分的两种设计。该图报告了5个数据集下的 ZP评估结果。

表 3.3: **三种区域划分下的 ZP方差：**空间偏差是检测器的外在表现，而空间均衡性则对应于给定的区域划分。

区域划分	VOC	COCO	人脸口罩	水果	安全帽
5 个环形区域	53.6	16.9	13.1	56.2	3.0
沿 x 轴的 5 个区域	32.3	7.2	11.2	13.7	6.4
沿 y 轴的 5 个区域	46.7	14.0	39.6	20.8	30.5

到39.6和30.5，而在这两种情况下，水果数据集的 ZP方差都显著降低。

*启示：*区域评估提供了一个新的视角，揭示了目标检测器的局限性。可以看出，空间偏差也是目标检测器的自然特征，并且对于任意区域划分，它们很难实现完美的空间均衡。以上结果表明，ZP 方差是与区域划分相关的集合函数。环形区域划分主要考虑内部区域和外部区域之间检测能力的平衡，这在实践中是一个不错的选择，因为中心化摄影师偏差在视觉数据集中普遍存在。然而，应该注意的是，区域划分是灵活的，并且能够根据应用场景定制为任何形



(a) GFocal (b) Cascade R-CNN (c) Deformable DETR

图 3.6: 具有各种目标尺度范围的平均 ZP。可以看出，对于每个目标尺度范围 r，三种目标检测器的空间偏差都很显著。

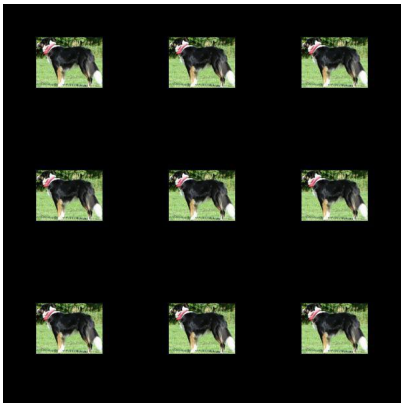
状。

3.4 空间偏差的主要来源

本节将进行探索性实验，以阐明空间偏差的主要来源。本节采用了三种具有代表性的目标检测器。第一个是流行的单阶段密集目标检测器 GFocal [115]。第二个是经典的多阶段由密集到稀疏的目标检测器 Cascade R-CNN [102]。第三个是稀疏目标检测器 Deformable DETR [118]。

3.4.1 目标尺度研究

从评估区域的定义（公式 3.1）中，读者可能会问目标尺度是否在中心区域性能中起关键作用。在本实验中，采用环形区域划分。如果目标的中心点坐标位于 $z^{i,j}$ 中，则该目标属于 $z^{i,j}$。为了消除目标尺度的影响，本文将区域评估过程限制在具有相似尺度的目标中。对于每个目标尺度范围 r，本文分别选择所有区域在 $[0,r^2],[r^2,(2r)^2],\cdots,[(kr)^2,\infty]$ 范围内的真实框，其中尺度的最大端点设置为 $kr=256$，并且 $r\in\{4,8,16,32,64,128,\infty\}$. 然后，本文计算所有尺度上ZP的平均值。如图 3.6所示，无论目标尺度范围选择得多小，空间偏差都非常显著。 $ZP^{4,5}$ 是最高的，相比之下，$ZP^{0,1}$ 是最低的。当评估区域更靠近图像边界时，性能下降幅度更大。可以看出，内部区域和外部区域之间的性能差距始终很大，ZP差距超过10。这表明中心化的空间偏差可能不是源于目标尺度因素。为简单起见，本文在以下实验中对区域评估采用所有尺度。



(a) 九宫格数据集

32.7 z^{11}	31.1 z^{12}	31.1 z^{13}
29.6 z^{21}	27.8 z^{22}	28.5 z^{23}
30.6 z^{31}	28.9 z^{32}	29.3 z^{33}

(b) 9个区域上的 ZP

图 3.7: (a) 九宫格数据集是通过将测试集的所有目标规则地放置在 600×600 的黑色图像上而构建的。(b) GFocal的区域评估（3×3 网格）。

3.4.2 目标绝对空间位置研究

由于区域性能表现出明显的中心化趋势，因此一个直接的推测是空间偏差与目标的绝对空间位置有关。如果目标被移动到中心区域，则可能会被更好地检测到，反之，如果在边界区域，则会更差。为了分析目标的空间位置是否在其检测质量中起关键作用，本文构建了九宫格数据集，用于检测在一张 600×600 纯黑色图像上被规则放置的目标。实验基于以下3个步骤： (1) 首先从PASCAL VOC 2007测试集 [36]中裁剪目标，总共14976个目标。 (2) 所有目标都被缩放到固定大小，并以 3×3 网格方式放置。见图 3.7(a)。 (3) 为了衡量每个网格的检测质量，评估区域被定义为相同的 3×3 区域，表示为 z^{ij}，$i, j \in \{1,2,3\}$。从图 3.7(b)可以看出，检测器在中心区域 z^{22} 中表现并非最佳，甚至是最差的。这种现象与 [167]中的先前观察结果有些不符，后者分析了平移100张图像的效果，并得出结论，当目标靠近图像边界时，检测器的性能可能会下降。然而，当样本数量增加到超过14000时，本文观察到，将目标移动到中心区域在统计学上不会导致检测质量的提高。因此，检测性能的中心化趋势与目标的绝对位置之间的相关性不太明显。

讨论：图 3.6的结论是中心目标比边界目标更容易被检测到，并且这与目标尺度无关，而图 3.7的结论是，目标的绝对位置（通过目标平移）与形成中心化区域性能无关。那么什么才是真正决定目标检测器空间偏差的因素？

3.4.3 区域之间的目标数据模式

本小节旨在研究目标检测器中心化空间偏差的来源，其背后的灵感是，中心化的空间偏差可能来自于区域之间目标数据模式的差异。如果一个目标是从中心区域风格的数据分布中采样的，则可以更好地检测到该目标，而如果它是从边界区域风格的数据分布中采样的，则检测会更差。本文对目标数据模式的表示非常通用 [270–272]，并且只需要对检测流程进行少量适应的修改，仅需要1）输入图像 I，2）所有真实框 G，3）来自预训练目标检测器的特征提取器 $f: I \to \mathbb{R}_{M\times H\times W}$. 在推理过程中，本文使用真实框从 $f(I)$ 中裁剪目标特征，然后沿空间维度对特征值取平均。每个目标都由一个 M 维特征向量 g 表示，该向量编码了高维空间中的目标数据模式。在本实验中，区域数设置为2。本文将中心区域表示为 z^{in}，它是一个矩形区域，左上角坐标为 $p=(0.25W, 0.25H)$，右下角坐标为 $q=(0.75W, 0.75H)$，其中 W 和 H 是输入图像的宽和高。除此之外的其余部分设置为边界区域，表示为 z^{out}。区域之间目标数据模式的差异表示为：

$$\mathcal{E}((G_1,u),(G_2,v)) = \frac{1}{KCM}\sum_{k=1}^{K}\sum_{c=1}^{C}\sum_{m=1}^{M}\mathbb{1}_k||\bar{g}_{m,c}^{u,G_1} - \bar{g}_{m,c}^{v,G_2}||, \tag{3.3}$$

其中 $\bar{g}$ 表示特征表示中心，$u,v \in \{z^{in}, z^{out}\}$ 表示从中心区域或边界区域采样的目标，以及 $G_1, G_2 \in \{G_{train}, G_{test}\}$ 表示从训练集或测试集采样的目标。误差针对每个类别分别计算，然后取平均值。C 是类别的总数。此外本文还引入一个指示函数 $\mathbb{1}_k$，用以消除目标尺度的影响。当目标尺度在范围 $R = \{[((k-1)r)^2, (kr)^2]\} \bigcup \{[((K-1)r)^2, \infty]\}$, $k \in \{0,1,\cdots,K-1\}$ 的其中之一时，$\mathbb{1}_k$ 为1，否则为0。简而言之，$\mathcal{E}$ 测量四个集合的特征表示中心之间的距离，即来自训练集的中心区域目标、训练集的边界区域目标、测试集的中心区域目标、测试集的边界区域目标。

结果在图 3.8中报告。对于 VOC数据集，训练集是 VOC 2007 trainval，测试集是 VOC 2007 test。对于 COCO数据集，训练集是 COCO train2017，测试集是 COCO val2017。可以看出，从同一区域采样的目标比从不同区域采样的目标具有明显更低的差异。具体而言，测试集的中心区域目标与训练集的中心区域目标更相似，而测试集的边界区域目标与训练集的边界区域目标更相似。这表明目标数据模式在不同区域之间实际上是不同的，并且神经网络能够捕捉到这种差异。如图 3.9(a) 所示，区域性能在 VOC 2007 trainval集上是中心化的，因此它自然地在测试集上继承了相同的趋势。更耐人寻味的是，本文进一步可视化了

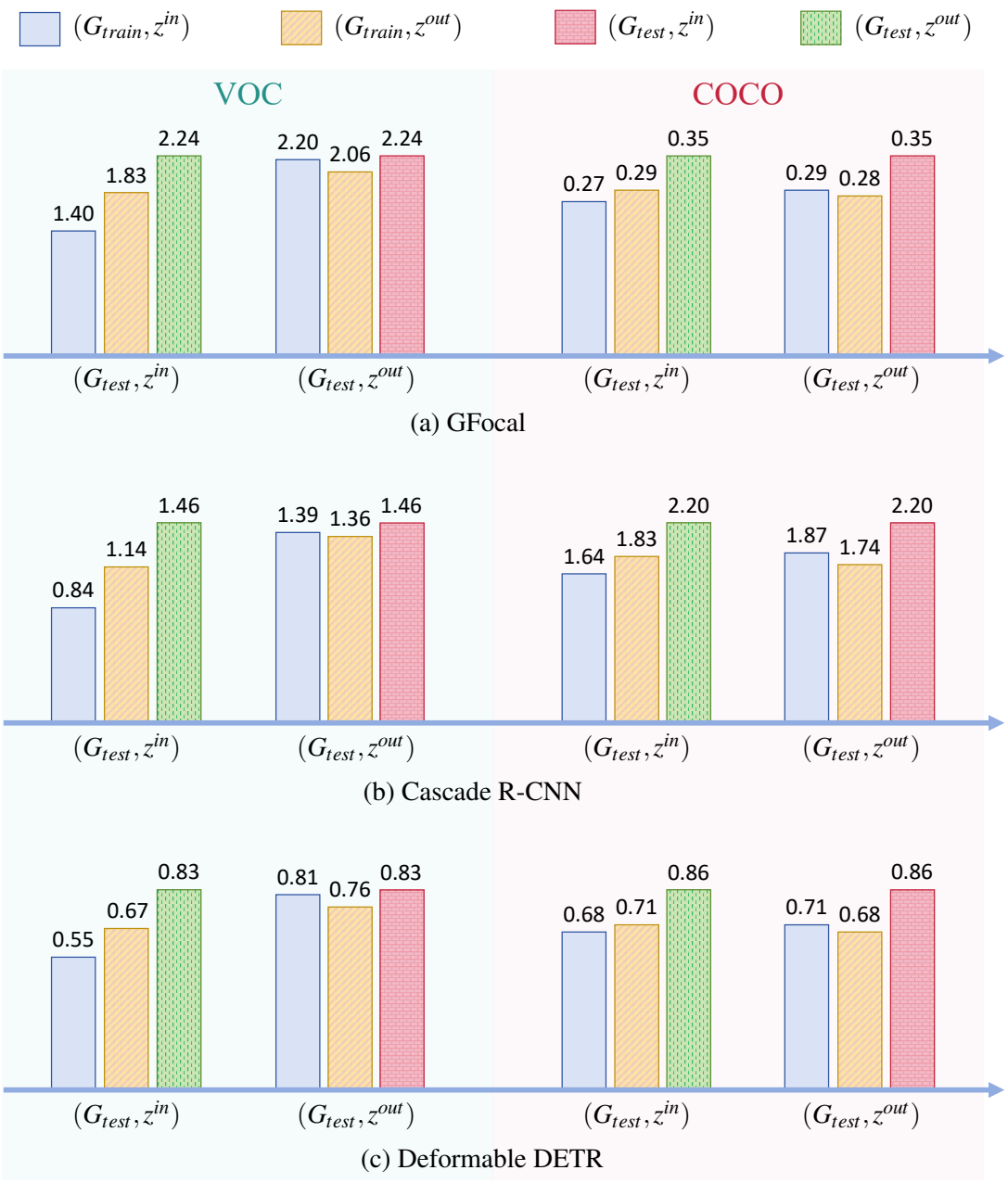


图 3.8: 特征表示中心之间的平均误差 $\mathcal{E}((G_1,u),(G_2,v))$。例如，蓝色条表示 $\mathcal{E}((G_{test},z^{in}),(G_{train},z^{in}))$。从同一区域采样的目标比从不同区域采样的目标具有显著更低的差异。

图 3.9(b-f) 中九宫格数据集上的检测性能，其中中心目标和边界目标是分开的。可以看出，无论我们将中心目标放置在哪里，检测器始终可以在检测中心目标方面表现更好。注意到，这种现象适用于所有5个数据集，包括 PASCAL VOC、MS COCO和其他3个应用数据集（口罩、水果、安全帽）。

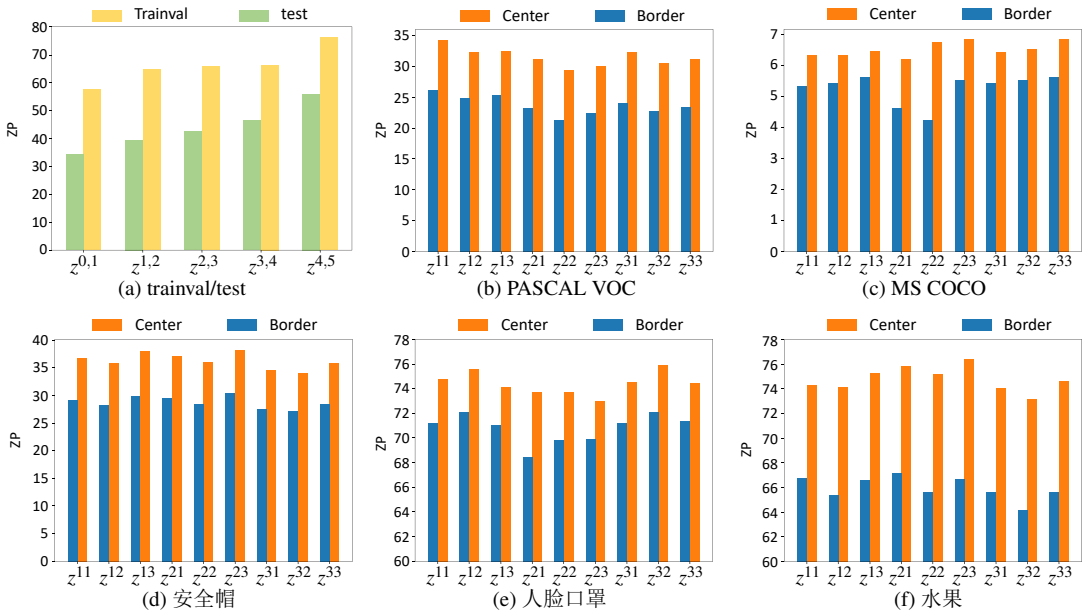


图 3.9: (a) VOC 2007 trainval集和 test集上的5区域评估。(b-f) 在九宫格数据集上，分别对中心目标和边界目标进行区域评估。结果表明，无论我们将中心目标放置在哪里，检测器始终可以在检测中心目标方面表现更好。

以上结果证实了本文的直觉，即如果目标是从中心区域风格的数据分布中采样的，则可以更好地检测到该目标，而如果目标是从边界区域风格的数据分布中采样的，则检测会更差。这表明，当我们人类拍照时，区域之间目标数据模式总是存在差异，而这种差异是人类难以察觉的。当镜头聚焦于目标最有可能出现的感兴趣区域时，就不可避免地会导致边界区域中目标的采样频率降低，从而导致次优的检测性能。

3.5 空间失衡问题

至此，本文已经展示了空间偏差的存在和主要来源。边界区域的次优性能阻碍了检测应用的鲁棒性。在本节中，将介绍目标检测的一个新的研究问题，空间失衡问题。

问题定义

设 S 为一个区域划分，m^S 表示一系列区域指标，$\sigma:\mathbb{R}^n\to\mathbb{R}$ 表示方差计算，则空间失衡问题被定义为最小化区域指标的方差：

$$\min_{\Theta}\sigma(m^S|\Theta), \tag{3.4}$$

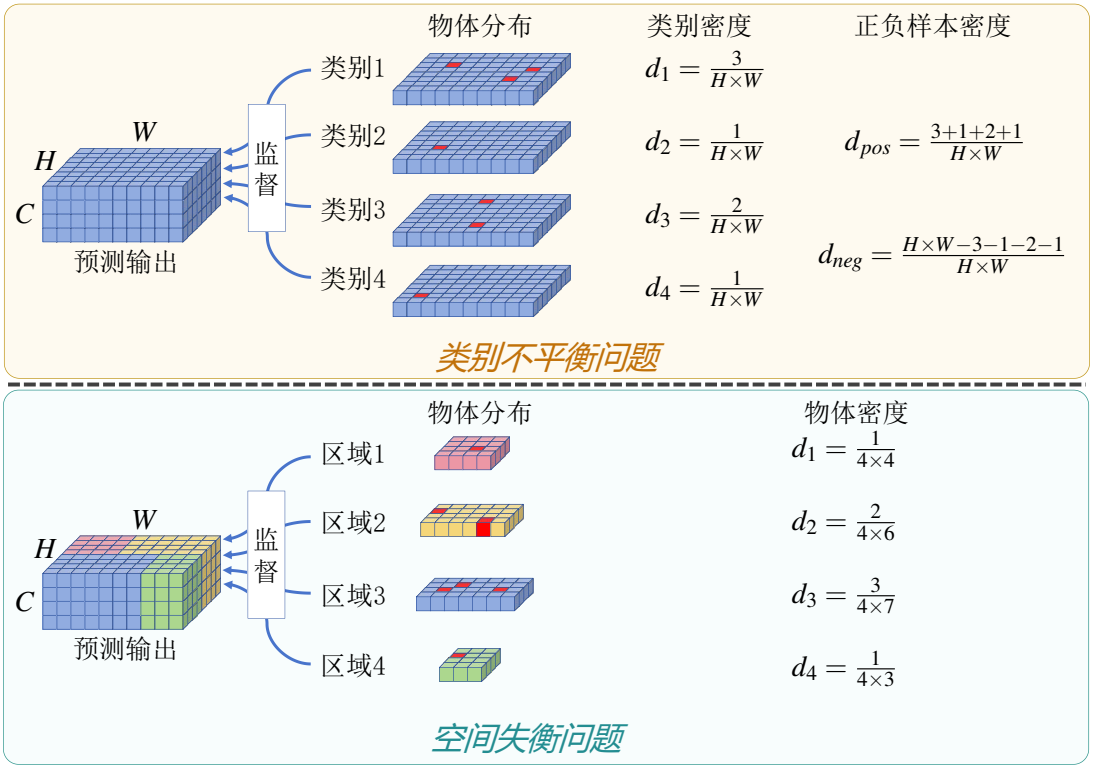


图 3.10: 类别不平衡问题和空间不均衡问题之间关系的说明。有 7 个 4 个类别的目标，用红色立方体表示。评估区域设置为 4 个区域，颜色分别为粉色、黄色、蓝色和绿色。本文简化了讨论，即每个目标仅包含 1 个正样本，多个正样本的情况类似。在左图中，类别密度是每个类别的目标数量与预测图大小的比率，而在右图中，目标密度是每个区域的区域目标数量与区域大小的比率。空间不均衡问题在形式上等同于类别不平衡问题。

其中 Θ 是目标检测器的网络参数集。促进空间均衡的总体目标主要取决于使用哪种区域划分，而这也取决于应用场景。因此，对于不同的应用场景，用户可以自定义区域划分。

讨论： 空间失衡问题在形式上等价于类别不平衡问题。

若将区域 z^i 中的目标表示为 X^i_{obj}，则给定区域 z^i 的目标密度可以表示为 $d_i = |X^i_{obj}|/|z^i|$。直观上，较高的密度表示更多的正样本，从而在区域上产生更大的梯度流。这类似于类别不平衡问题，后者在类别之间具有长尾分布，如图 3.10所示。具体而言，分类分支预测类别分数，这是一个 $C \times H \times W$ 的张量。第 c 个类别的密度表示为比值 $d_c = |X_c|/(H \times W)$，这里 X_c 是第 c 个类别的目标。由于 $H \times W$ 是一个常数，因此等价于将一个类别的所有样本放置在面积相同的

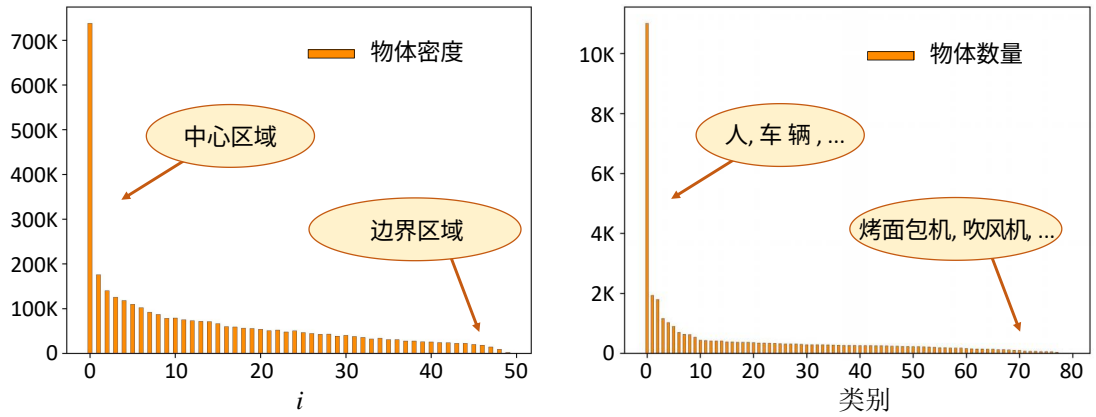


图 3.11: **左图：**COCO val2017上50个区域的目标密度。这些区域被中心定义为 $z^{i,i+1}$，$i=0,1,\cdots,49$。**右图：**COCO val2017上类别的长尾分布。空间失衡问题与类别不平衡问题具有相似的特征。

区域中。然后，每个类别都有一个用于模型学习的 $H\times W$ 的单个区域，并且类别之间是不相交的。从图 3.11中可以看出，这两个问题都服从长尾分布。由于这一事实，区域性能也可能与区域中的监督信号强度相关。一种简单的验证方法是增大或缩小区域的监督信号强度，以使网络达到新的收敛状态。这里，本文将一个新的参数 β 插入到标签分配算法 ATSS [114]中。如果锚点的 IoU大于 $\alpha_{pos}+\beta*\mathbb{1}_z$，则该锚点被分配为正样本，其中 α_{pos} 是正样本 IoU阈值。如果此锚点的中心点位于区域 z 中，$\mathbb{1}_z$ 则为1，否则为0。因此，当 β 增加时，正样本的数量 $|X^i_{pos}|$ 会相应减少。

本文在图 3.12中可视化了相对 ZP，其中所有 ZP都减去了 $\text{ZP}^{0,1}$。可以看出，如果通过减少边界区域中正样本的数量来削弱监督信号，则中心化空间偏差会进一步加剧。相反，如果在中心区域削弱监督信号，甚至可以实现反中心化的空间偏差。这表明监督信号强度确实对区域性能有影响。鉴于以上分析，本文最终讨论了一种可能的解决方案，用于解决环形区域划分下的空间失衡问题。

3.6 空间均衡学习

大多数现有的目标检测研究都侧重于追求图像级的更高检测性能，而忽略了区域级的优化，从而导致检测器出现严重的空间失衡问题。在本小节中，将介绍一种可能的解决方案，称为空间均衡学习，作为缓解空间失衡问题的开端。

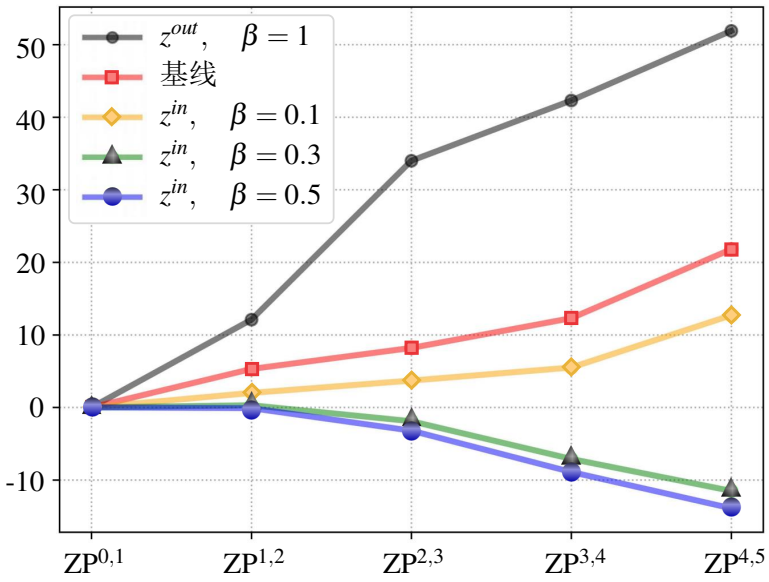


图 3.12: ZP相对于 $ZP^{0,1}$。如果边界区域的监督信号强度降低，则区域性能可以进一步极端中心化；反之，如果中心区域的监督信号强度降低，则区域性能可以是反中心化的。基线模型为 GFocal [115]原始模型。

首先介绍空间权重，它通过下式将锚点坐标 (x^a, y^a) 映射到标量 $\alpha(x^a, y^a)$：

$$\alpha(x,y) = 2\max\left\{||x-\frac{W}{2}||_1\frac{1}{W}, ||y-\frac{H}{2}||_1\frac{1}{H}\right\} \in [0,1], \tag{3.5}$$

其中 W 和 H 是图像的宽和高。空间权重可以很容易地插入到现有的检测流程中，只需进行少量修改。原理简单且具有多种选择。在这里，本文提供以下两种实现方式。

(1) 空间均衡标签分配

在这种方法中，关键思想是在制定标签分配的判定规则时，将空间权重视为一个额外的约束项。由于大多数标签分配算法都有其自身复杂的实现方式，因此为简洁起见，在下文中，本文提供了经典标签分配算法 ATSS [114]的具体应用描述。给定正样本 IoU阈值 t，该阈值是通过考虑目标的统计特性计算得到的。 ATSS准则遵循与 max-IoU分配 [96, 104, 106]相同的规则，即 $\text{IoU}(\boldsymbol{B}^a, \boldsymbol{B}^{gt}) \geqslant t$，其中 $\boldsymbol{B}^a$ 和 $\boldsymbol{B}^{gt}$ 分别表示预设的锚框和真实框。空间均衡标签分配（Spatial Equilibrium Label Assignment，SELA）过程表达为：

$$\text{IoU}(\boldsymbol{B}^a, \boldsymbol{B}^{gt}) \geqslant t - \gamma\alpha(x^a, y^a), \tag{3.6}$$

其中 $\gamma \geqslant 0$ 是一个超参数。可以看出，SELA放宽了图像边界附近目标的正样本

选择条件。因此，将有更多的锚点被选中作为它们的正样本。请注意，上述应用实际上是一种基于频率的方法，类似于为类别不平衡问题所提出的许多分类重平衡采样策略 [147, 148]，这类方法通常需要平衡头部与尾部类别的分类性能。

(2) 空间均衡损失

在这种方法中，本文采用代价敏感学习方法，将空间权重项 $1+\gamma\alpha(x^a,y^a)$ 作为分类和边界框回归损失的附加权重因子：

$$\mathcal{L} = (1+\gamma\alpha(x^a,y^a))\mathcal{L}. \tag{3.7}$$

这里 $\mathcal{L}$ 可以为分类损失 $\mathcal{L}_{\text{cls}}$ 或边界框回归损失 $\mathcal{L}_{\text{reg}}$。如此一来，模型训练时将在边界区域产生更大的梯度流，从而使网络更加关注边界目标。

3.7 空间均衡的探究实验

最后，本节提供关于空间均衡学习的探究实验与评估。消融实验使用GFocal [115]进行，并且本文默认采用章节 3.6中介绍的空间均衡学习中的第一种方法，即空间均衡标签分配 (SELA)。

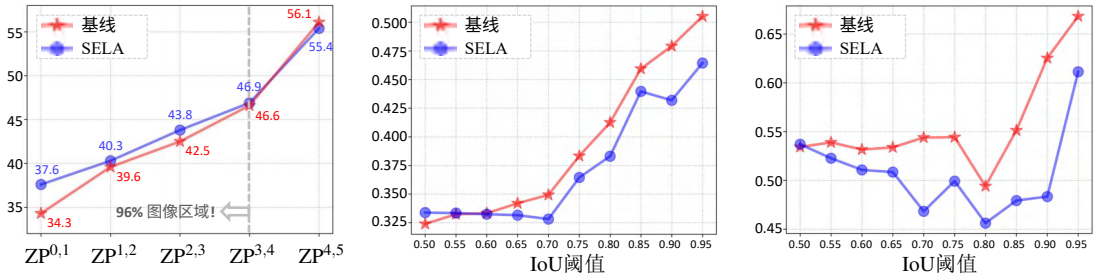
超参数 γ

回顾一下，SELA的实现仅涉及到公式 3.6中的一个新超参数 γ，用于控制空间权重的大小。较大的 γ 会为图像边界附近的目标增加更多的正样本。如表 3.4所示，SELA可以为 γ 的所有选择实现稳定的空间均衡改进（较低的方差）。过大的 γ，例如0.4，将为所有区域增加更多的正样本，导致性能下降。因此，本文将 PASCAL VOC的 γ 设置为0.2。可以看出，SELA可以显著提高外部区域（例如，$\text{ZP}^{0,1}$、$\text{ZP}^{1,2}$、$\text{ZP}^{2,3}$ 和 $\text{ZP}^{3,4}$）的检测性能。如图 3.13(a) 所示，虽然中心区域 $z^{4,5}$ 的性能略有下降，但 ZP的改善在边界区域仍非常显著，而边界区域占据了总张图像的96%的面积。这对于监控系统和自动驾驶中的安全应用尤为重要，因为目标可能出现在任何地方。边界区域的性能在鲁棒性检测中起着重要作用。在实践中，本文将所有其它数据集的 γ 设置为0.1，但应注意，对于不同的应用场景，可能存在更好的 γ。

空间权重

表 3.4: **SELA中超参数 γ 的评估：**该表报告了5个区域精度 (ZP)、ZP的方差和传统指标 AP。$\gamma = 0$ 表示基线 GFocal。方差越低，空间均衡性越好。（数据集：VOC 07+12）

γ	AP	方差	$ZP^{0,1}$	$ZP^{1,2}$	$ZP^{2,3}$	$ZP^{3,4}$	$ZP^{4,5}$
0	52.2	53.6	34.3	39.6	42.5	46.6	**56.1**
0.1	52.5	44.5	35.9	40.6	42.1	46.6	55.6
0.2	**52.8**	**37.7**	**37.6**	40.3	**43.8**	**46.9**	55.4
0.3	**52.8**	37.3	37.4	**41.5**	43.6	**46.9**	55.6
0.4	52.0	46.3	35.0	38.9	42.6	46.6	54.8



(a) SELA 的效果 (b) PCC (c) SCC

图 3.13: (a) ZP与区域的关系。 (b) mZP与目标分布（中心计数）之间针对 IoU阈值的 Pearson相关系数(PCC)。(c) mZP与目标分布之间针对 IoU阈值的 Spearman相关系数(SCC)。本文的 SELA可以在大多数 IoU阈值下大幅降低这些相关性，表明空间均衡性更好。基线模型为 GFocal。结果在 VOC 07+12上报告。

人们可能想知道，如果本文直接放宽正样本的选择条件而不考虑其所处的空间位置，性能会如何变化。在这里，本文进行实验以研究空间权重的影响。定量结果在表 3.5中报告。如果空间权重设置为常数1，则意味着直接将正样本 IoU阈值 t 降低为 $\text{IoU}(\boldsymbol{B}^a, \boldsymbol{B}^{gt}) \geqslant t - \gamma$，这意味着将选择更多的正样本而没有空间位置的区分。可以看出，尽管性能有所提高，但 ZP方差仍然很大。这表明从正样本 IoU阈值中减去一个常数不能显著改变采样频率，因为在中心区域也会生成更多正样本。相比之下，SELA可以显著降低 ZP方差，并实现更好的空间均衡性。

各种数据集上的 SELA

接下来，表 3.6展示了有希望的结果，即本文的 SELA可以在更多应用场景下为目标检测器实现更好的空间均衡性。特别是，本文方法大幅降低了 ZP方差。例如，在 PASCAL VOC、MS COCO和口罩/水果检测方面，本文方法

表 3.5: **空间权重的分析：** 该表报告了5个区域精度(ZP)、ZP的方差和传统指标AP。 $\gamma = 0.2$。方差越低，空间均衡性越好。（数据集：VOC 07+12）

权重	AP	方差	$ZP^{0,1}$	$ZP^{1,2}$	$ZP^{2,3}$	$ZP^{3,4}$	$ZP^{4,5}$
0	52.2	53.6	34.3	39.6	42.5	46.6	56.1
1	**52.8**	48.3	35.7	40.2	43.3	**47.1**	**56.2**
$\alpha(x^a, y^a)$	**52.8**	**37.7**	**37.6**	**40.3**	**43.8**	46.9	55.4

表 3.6: 空间均衡学习在 PASCAL VOC 07+12、MS COCO 2017和 3个应用数据集（包括人脸口罩检测、水果检测和安全帽检测）上的定量结果。该表报告了5个区域精度(ZP)、ZP 的方差和传统指标 AP。

数据集	SELA	AP	方差	$ZP^{0,1}$	$ZP^{1,2}$	$ZP^{2,3}$	$ZP^{3,4}$	$ZP^{4,5}$
VOC 07+12		52.2	53.6	34.3	39.6	42.5	46.6	56.1
	✓	52.8	37.7	37.6	40.3	43.8	46.9	55.4
COCO 2017		40.1	16.9	31.1	37.5	39.4	38.5	43.8
	✓	40.3	14.4	31.2	37.7	39.5	38.3	42.9
人脸口罩		71.3	13.1	60.4	67.1	69.0	68.8	70.9
	✓	71.6	12.1	60.6	68.0	69.5	69.3	69.8
水果		76.6	56.2	60.8	69.9	71.2	75.3	83.8
	✓	77.0	33.6	65.7	69.8	72.0	76.2	82.7
安全帽		49.7	3.0	45.9	47.9	50.3	50.6	47.8
	✓	49.9	3.1	45.9	48.5	50.5	50.6	47.9

SELA成功地将 ZP的方差降低了15.9、2.5、1.0 和22.6，而 AP仍被保持或有微小提升。这表明 SELA可以提高多种应用场景的空间均衡性，而不会牺牲 AP。

各种检测模型上的 SELA

本节将验证空间均衡学习在各种检测模型上的有效性。首先是3个不同的主干网络。 表 3.7的结果表明 SELA可以显著提高所有3个主干网络的空间均衡性（降低 ZP方差）。此外，本节还将空间均衡学习结合到更多的目标检测器中，如 DW [132]、DDOD [273]和 DETR类型检测器 DINO [124]，来检验空间均衡学习的通用性。在这里，本文采用空间均衡损失（SE损失），其扩大了图像边界区域附近目标的训练损失。 表 3.8报告了4个目标检测器的 SE损失的定量结果。可以看到本文方法可以显著降低4个检测器的 ZP方差，这表明实现了更好的空间均衡性。这显示了本文所提出的空间均衡学习的通用性，其无需任何花里胡哨的技术下即可提高检测器空间鲁棒性。

表 3.7: **各种主干网络的 SELA评估：**该表报告了5个区域精度(ZP)、ZP方差和传统指标 AP。**X**: ResNeXt [267]。（数据集： VOC 07+12）

模型	SELA	AP	方差	$ZP^{0,1}$	$ZP^{1,2}$	$ZP^{2,3}$	$ZP^{3,4}$	$ZP^{4,5}$
ResNet-18		52.2	53.6	34.3	39.6	42.5	46.6	56.1
	✓	52.8	37.7	37.6	40.3	43.8	46.9	55.4
ResNet-50		56.1	41.5	40.9	44.6	46.7	51.0	59.7
	✓	56.2	32.2	43.3	44.6	47.3	50.4	59.2
X-101-32x4d-DCN		64.0	37.1	48.7	53.1	55.0	58.0	66.9
	✓	64.3	31.0	50.2	54.1	55.9	57.7	66.9

表 3.8: **各种检测器的 SELA评估：**这里采用了基于代价敏感学习的方法，即SE损失。该表报告了5个区域精度(ZP)、ZP方差和传统指标 AP。（数据集： VOC 07+12）

检测器	SE 损失	AP	方差	$ZP^{0,1}$	$ZP^{1,2}$	$ZP^{2,3}$	$ZP^{3,4}$	$ZP^{4,5}$
GFocal [115]		52.2	53.6	34.3	39.6	42.5	46.6	56.1
	✓	52.5	41.6	37.1	40.6	42.9	46.5	56.0
DW [132]		51.8	32.6	38.4	39.9	43.3	45.7	54.6
	✓	52.7	25.9	39.8	41.2	44.4	46.8	54.2
DDOD [273]		51.1	22.6	38.4	40.0	42.2	45.2	51.9
	✓	51.5	20.8	40.9	40.1	42.6	45.8	52.7
DINO [124]		61.5	47.6	47.1	48.4	53.0	57.1	66.2
	✓	61.7	46.7	47.4	48.5	53.4	57.1	66.3

本文还注意到，与基于 CNN的目标检测器相比，本文方法在 DINO上产生了较为轻微的空间均衡改进。这可能归因于 DETR类型检测器和其它检测器之间不同的优化过程。 DETR类型检测器中的正样本数量非常有限，因为它们使用一对一的匈牙利匹配，而密集目标检测器中正样本则丰富得多，因为它们采用一对多的标签分配。因此，本文的 SE损失更有助于缓解密集目标检测器上监督信号强度不平衡的问题。这意味着改善 DETR类型检测器的空间均衡性可能更具挑战性。本文希望这项工作可以启发更多解决方案，以解决未来 DETR类型检测器的空间失衡问题。

与目标分布的相关性

本节进一步提供了区域指标与目标分布之间的相关性。这里定义了更精细的区域划分，与用于计数目标中心的区域划分相同，即 11×11 个方形区域（见

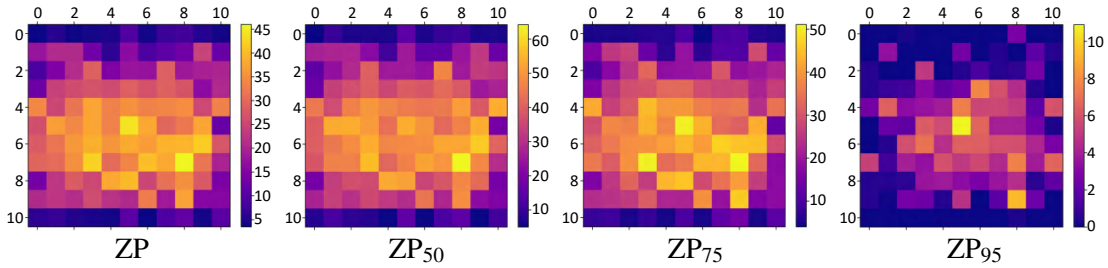


图 3.14: 在 11×11 个方形区域上进行区域评估。模型是 GFocal。结果在 VOC 07+12上报告。

图 3.4）。然后，逐个评估121个区域的检测性能，并在图 3.14中绘制了121个区域的 ZP。可以看出，ZP分布与目标分布（图. 3.4）高度相似，即相同的中心化趋势。为了研究区域指标与目标分布之间的相关性，本文进一步计算了 ZP与测试集目标分布之间的 Pearson相关系数(PCC)和 Spearman相关系数(SCC)。如图 3.13(b)和图 3.13(c)所示，本文对空间偏差有了以下深刻的反思。首先注意到，图 3.13(b)中的所有 PCC > 0.3，这表明检测性能与目标分布呈中度线性相关。注意到，PCC仅反映两个给定向量的线性相关性，而当它们呈曲线相关时，使用 PCC可能会失效。在图 3.13(c)中，Spearman相关性反映了 ZP和目标分布之间更高的排序相关性，所有 SCC > 0.45。这说明检测性能与目标分布具有中等到高度的相关性。而本文的 SELA大大降低了这些相关性，表明与目标分布的相关性较低，空间均衡性更好。

检测可视化

图 3.15中可视化了 SELA的检测结果。本文方法可以提高边界区域的检测性能。本文认为，进一步探索空间均衡性对于鲁棒的检测应用显然是值得且重要的。

其他实现空间均衡的尝试

回顾本文在章节 3.4中所讨论的，目标尺度和目标的绝对位置几乎不影响空间偏差。区域之间目标数据模式的差异在空间偏差中起着重要作用。在这里，本文研究更多可能的因素，看看这些因素会对空间均衡性产生什么变化。首先第一个所考虑的因素是填充操作。本文遵循 Kayhan, O. S.等人的研究工作 [159]，将填充方式设置为 full-conv，正如他们所证明的那样，它是平移不变的。本文使用 full-conv替换了检测头网络的所有卷积核。第二个考虑的因素是图像边界

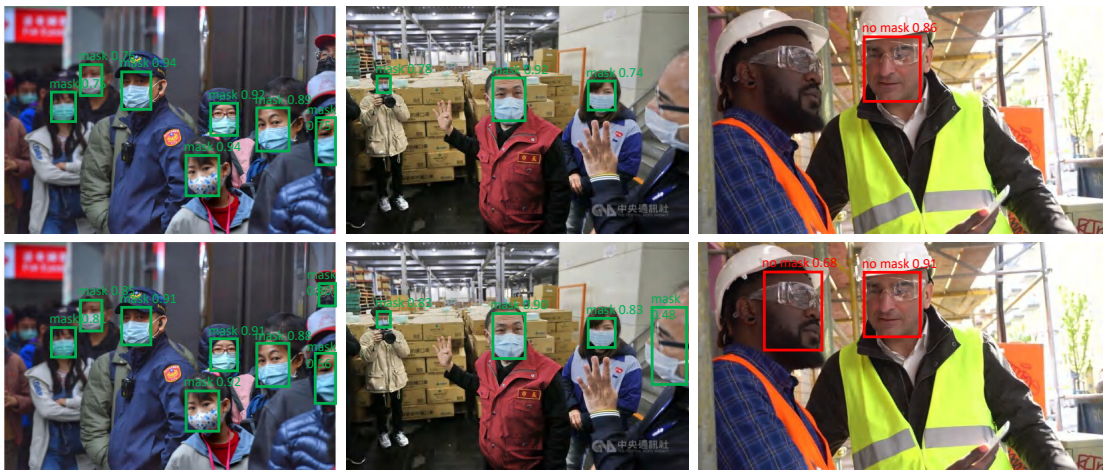


图 3.15: GFocal（第一行）和 GFocal + SELA（第二行）的检测结果图示。本文方法提升了边界区域的检测性能。放大以获得更好的视图。

表 3.9: **空间均衡的3个潜在因素的评估：**(1) 本文在检测头网络中使用 full-conv [159]填充操作；(2) 本文移除了超出有效图像边界的过大锚框；(3) 本文将训练图像分辨率设置为 640×640。该表报告了5个区域精度(ZP)、ZP方差和传统指标 AP。结果在 COCO val2017上的 GFocal [115]上报告。

修改方法	AP	方差	$ZP^{0,1}$	$ZP^{1,2}$	$ZP^{2,3}$	$ZP^{3,4}$	$ZP^{4,5}$
基线	40.1	16.9	31.1	37.5	39.4	38.5	43.8
(1)	38.6	12.4	30.4	36.0	37.5	37.8	41.2
(2)	38.5	16.7	28.8	35.8	37.8	37.2	41.3
(3)	36.9	18.8	26.7	33.6	36.2	34.9	39.9

附近过大尺寸锚框的影响。这里基线模型的默认设置保留了所有锚框。作为对照，本文移除了所有边界超出有效图像范围的锚框。第三个考虑的因素是图像分辨率。基线模型的训练默认分辨率为 1333×800，本文训练了一个分辨率较小的模型，例如 640×640。结果在表 3.9中报告。可以看出，以上三个因素都导致了 AP的显著下降。填充操作 full-conv可以降低 ZP方差，但对检测精度的提高没有帮助。此外，通过移除越界锚框或者设置不同的训练图像分辨率也无法获得更好的空间均衡性，这是因为区域之间的监督信号仍然不平衡。在未来，找到一种既能缓解空间失衡问题又不会导致性能下降的解决方案将具有挑战性。

3.8 本章小结

本章提出了区域评估，以揭示现代目标检测器中空间偏差的存在和离散幅

度。本文发现，空间偏差与目标尺度和目标的绝对位置的相关性较小，而与区域之间目标数据模式的差距密切相关。基于对空间偏差起源的深入研究，本章最终提出了空间失衡问题，旨在实现跨区域的鲁棒检测。作为缓解此问题的开始，本章还展示了一条通往空间均衡目标检测的路径，即空间均衡学习。广泛的实验证明了空间偏差的存在和主要来源，这在各种现代检测器和数据集中普遍存在。

意义： 空间偏差是目标检测中的一个天然障碍，检测器通常在边界区域表现出性能下降，而边界区域占据了图像区域的很大一部分。虽然经典的 AP指标仍然被认为是主要的衡量标准，但它很难揭示空间偏差，并且难以全面反映目标检测器的真实性能。最大化 AP指标并不能完全表明鲁棒检测，并且在所有区域都表现良好。区域评估补充了一系列区域指标，弥补了传统评估的缺点，并捕获了更多关于检测性能的信息。希望本文的研究能够启发社区重新思考目标检测器的评估，并激发对空间偏差以及空间失衡问题解决方案的进一步探索。

这本章的研究仍遗留下了一些挑战：

各种目标检测器中空间偏差的可解释性： 本章主要揭示了目标检测器中空间偏差的存在和离散幅度，而不同检测器表现差异很大的具体原因仍然扑朔迷离。神经网络架构设计、预训练数据、优化、训练策略，甚至超参数都可能在空间偏差中发挥作用。进一步探索以回答上述问题至关重要。

其他潜在因素对空间偏差的影响： 目前，本章指出了不平衡的目标分布与区域性能之间存在明显的关联。还有一些复杂而隐含的因素，例如图像模糊、目标遮挡、边界效应、噪声等，也可能导致空间偏差。然而，当前检测数据集几乎缺乏对上述因素的注释，这使得难以建立定量分析。

其他视觉任务的区域评估： 研究人员发现了一些线索，表明图像生成器可能会在图像边界附近生成失真内容 [164]。因此，空间偏差也可能存在于许多视觉任务中。本文的区域评估可能具有巨大的潜力来揭示空间偏差，无论是对于高级还是低级视觉任务。

第四章 定位蒸馏：目标检测的紧致表达技术

目标检测在通往高效能的路上，如何进行紧致表达一直以来都是学界的研究热点。知识蒸馏（KD）在目标检测中具有学习紧致模型的强大能力。本章首先指出了目标检测界对于蒸馏定位头的误区，即以往的 KD方法无法应用于定位头，导致定位知识传递效率低下。通过对定位知识蒸馏过程的重新定义，本章节提出了一种新颖的定位蒸馏（Localization Distillation，LD）方法，可以有效地将教师模型的定位知识转移到学生模型中。随后本章进一步提出选择性区域蒸馏法，通过引入有价值定位区域的概念，使得不同的蒸馏方法可以有选择性地在各自利好的区域上进行。结合这两个新组件，本章节首次表明 logit模仿可以胜过特征模仿，并且缺乏定位蒸馏是 logit模仿多年来表现不佳的一个关键原因。最后，基于定位蒸馏，本章节还将研究 logit模仿与特征模仿的优劣性。深入研究展示了 logit模仿的巨大潜力，其可以显著减轻定位的模糊性，学习鲁棒的特征表示，并在训练的早期阶段减轻训练困难。本章节的实验将主要在 MS COCO、PASCAL VOC和 DOTA基准测试上进行，实验结果表明本文的定位蒸馏方法可以在不降低推断速度的情况下实现显著的平均精度（AP）提升。

4.1 引言

作为一种模型压缩技术，知识蒸馏（KD） [205, 206]已成为一种有效的方法，用于学习紧致的模型以减轻计算负担。通过将大型教师网络捕获的泛化知识传递给小型学生网络，知识蒸馏对于提升小型学生网络性能的有效性已得到广泛验证 [205, 206, 274–277]。在目标检测中，关于知识蒸馏主要有三种流行的蒸馏流程，如图 4.1所示。首先是 logit模仿 [205]，也被称为分类蒸馏，最初是针对图像分类而设计的，其中蒸馏过程针对师生对的 logit进行操作。其次是特征模仿，受先驱工作 FitNets [206]的启发，旨在强制师生对之间的特征表示相一致。最后，伪边界框回归使用来自教师的预测边界框作为对学生的边界框预测分支的额外监督。

在这些方法中，最初的 logit模仿技术 [205]用于分类通常效率低下，因为它只传递分类知识，而忽视了定位知识蒸馏的重要性。因此，目标检测中现有

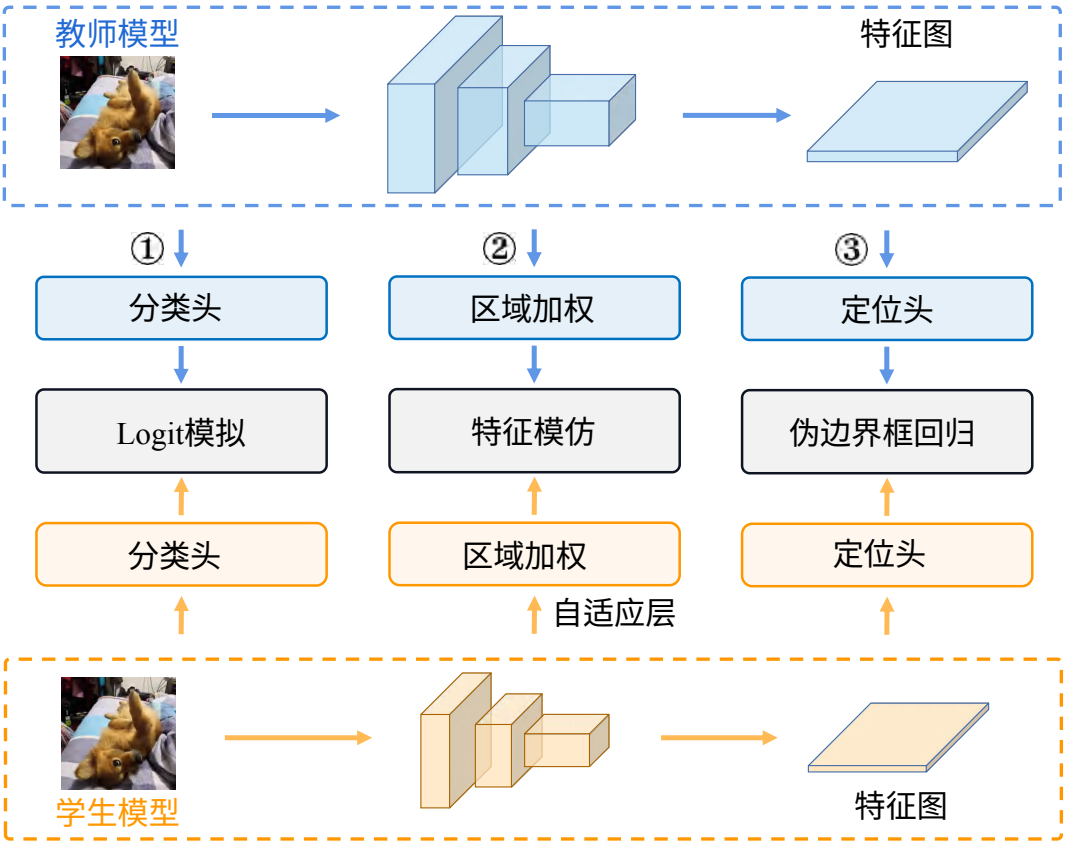


图 4.1: 目标检测中现有的知识蒸馏框架。 ① Logit模仿（分类蒸馏）：在 [205]中提出的分类蒸馏方法。 ② 特征模仿：最近的流行方法，基于不同的蒸馏区域蒸馏中间特征，通常需要自适应层来对齐学生的特征图的尺寸。 ③ 伪边界框回归：将教师的预测边界框视为额外的回归目标 [207, 212]。

的知识蒸馏方法主要侧重于特征模仿，并且表明蒸馏特征表示比蒸馏 logits更有优势 [209, 278, 279]。本文总结了这种现象的三个关键原因：首先，logit模仿的有效性在一定程度上取决于类别数量，而不同的应用场景中类别数量可能不同 [209]。第二，logit模仿只能应用于分类头部，无法蒸馏定位信息。第三，在多任务学习的框架下，特征模仿可以传递分类和定位的混合知识，从而有利于下游的分类和定位任务。

在本工作中，我们考察了目标检测知识蒸馏中先前提出的普遍观点，即分类蒸馏能否用于定位头？随后本章节检查特征模仿是否总是领先于 logits模仿？为此，本文首先提出了一种简单而有效的定位蒸馏（LD）方法，受到一个有趣的观察启发：教师生成的边界框分布 [115, 198]可以作为对学生检测器的强监督。边界框分布 [115, 198]最初设计用于建模真实边界框的分布，是解决定位

模糊性的高效方法，如图 4.2所示。 通过离散化的概率分布表示，定位器可以

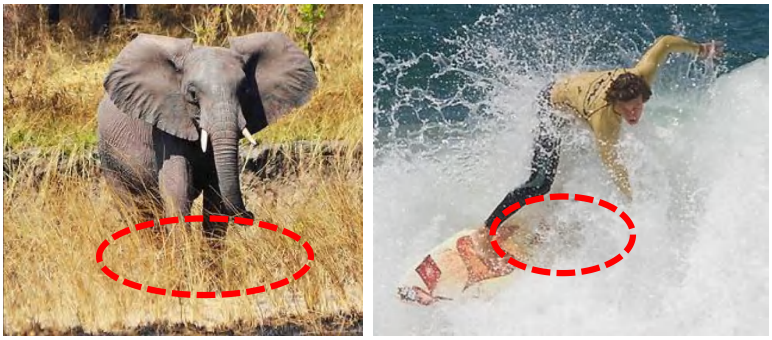


图 4.2: 边界框的定位模糊性：“大象”的下边界和“冲浪板”的右边界是模糊的

通过分布的平坦程度和锐度来反映定位的模糊性，而这在传统的边界框的狄拉克 δ 函数表示 [96, 101, 109, 110]中是不具备的。这使得本文的定位蒸馏方法能够从教师到学生更有效地传递更丰富的定位知识，而不是仅使用伪边界框回归（图 4.1中的右部分）。

将提出的定位蒸馏（LD）方法与分类蒸馏相结合，进而形成了一种统一的知识蒸馏方法，基于纯 logit模仿的框架，适用于分类分支和定位分支。由于 logit模仿使我们能够分别蒸馏分类知识和定位知识，本文发现这两个子任务对不同的蒸馏区域有不同的偏好。受这一观察启发，本文引入了有价值定位区域（VLR）的概念，并提出以选择性区域蒸馏的方式进行蒸馏。在实验部分，本文将展示在本文的蒸馏框架中使用 VLR的优势。

此外，本章节还将全面讨论 LD的技术细节，并详细阐述了 logit模仿和特征模仿的行为。有趣的是，本文观察到 logit模仿首次能够胜过特征模仿，这表明定位蒸馏的缺失实际上是 logit模仿在目标检测中多年表现不佳的关键原因。另一个观察是，本文发现 logit模仿有效的原因不是因为师生对之间特征表示的一致性被加强。恰恰相反，从师生特征 l_n 距离以及线性相关性的角度来看，学生的特征表示与教师的特征表示的差异被显著放大。本文还观察到，如果使用特征模仿训练学生模型，它倾向于在特征子空间中产生一个尖锐的 AP得分分布，并加剧了早期训练阶段的训练难度。

上述观察反映了 logits模仿相对于特征模仿的巨大潜力： 1）能够分别传递不同类型的知识； 2）学习更加鲁棒的特征表示； 3）减轻训练难度。本文的方法简单且易于实施，可以轻松应用于水平和旋转物体检测器中，以提高它们的性能，而无需引入任何推断开销。在 MS COCO上进行的大量实验证明，

本文的方法无需繁琐的设计，仅使用 ResNet-50-FPN骨干网络，在强基线模型 GFocal [115]的基础上，将 AP得分从40.1提升到42.1，将 AP_{75} 得分从43.1提升到45.6。本文使用 ResNeXt-101-32x4d-DCN骨干网络的最佳模型可以实现50.5 AP的单尺度测试，超越了在相同骨干、neck和测试设置下的所有现有检测器。

本章节的主要贡献包括以下四个方面：

- 本章节提出了一种新颖的定位蒸馏方法，极大地提高了目标检测中 logit模仿的蒸馏效率。
- 本章节对 logit模仿和特征模仿的行为进行了探索性实验和分析。这是首次揭示 logit模仿相对于特征模仿的巨大潜力的工作。
- 本章节提出了基于新引入的有价值的定位区域的选择性区域蒸馏方法，以更好地蒸馏学生检测器。
- 本章节将定位蒸馏扩展到旋转版本，使其可以应用于有向目标检测。

4.2 定位蒸馏方法描述

在目标检测的知识蒸馏流程中，输入图像被输入到两个目标检测器中，即学生检测器和冻结的教师检测器。蒸馏过程要求学生的输出模仿教师的输出。目标检测中有两种主流的知识蒸馏方法范式。

Logit模仿：最初用于图像分类 [205]，也可被称为分类 KD 其中学生模型可以通过模仿教师分类器的软输出来得到改进。令 $\boldsymbol{z}_S, \boldsymbol{z}_T \in \mathbb{R}^{W\times H\times C}$ 为学生和教师预测的 logits，其中 W 和 H 表示 logit图的输出尺寸，C 表示类别数量。然后，通过使用广义 SoftMax函数将这些 logits转换为概率分布 $\boldsymbol{p}_\tau$ 和 $\boldsymbol{q}_\tau$。知识蒸馏通过最小化以下损失来训练网络：

$$\mathcal{L} = \mathcal{L}_{CE} + \lambda \mathcal{L}_{KD} \tag{4.1}$$

$$= \mathcal{H}(p, g) + \lambda \mathcal{H}_(p_\tau, q_\tau), \tag{4.2}$$

其中，$\boldsymbol{p}$ 是预测的概率向量，$\boldsymbol{g} = \{0,1\}^n$ 是真实标签的 one-hot向量，$\mathcal{H}$ 表示交叉熵损失，λ 平衡了两个损失项。对于目标检测，可以在一些预定义的蒸馏区域 $\mathcal{R}$ 上进行蒸馏。

特征模仿：旨在通过模仿教师和学生之间的深层特征来传递知识 [206, 209]。数

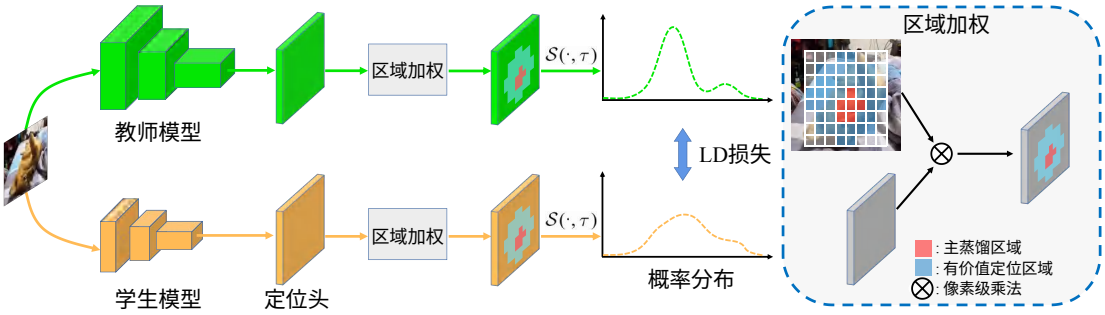


图 4.3: 定位蒸馏框架图

学上，特征模仿过程可以表示为：

$$\mathcal{L}_{FI} = \frac{1}{|\mathcal{R}|} \sum_{r \in \mathcal{R}} ||\tilde{M}_S(r) - M_T(r)||_2, \tag{4.3}$$

其中，$\mathcal{R}$ 表示模仿的区域，$|\cdot|$ 表示区域的基数。需注意的是，由于学生特征图与教师特征图的尺寸可能并不一致，则需要使用额外的自适应层将学生的特征图 M_S 的尺寸转换为与教师的特征图 M_T 相同的尺寸，即 $\tilde{M}_S, M_T \in \mathbb{R}^{W \times H \times D}$。

在章节 2.2中，本文介绍了边界框表示的 t,b,l,r 形式，即采样点到边界框的上、下、左、右四条边的距离 [101]。实际上，这种边界框的四元表示等价为每条边服从狄拉克 δ 分布，即只关注真实位置，无法模拟定位模糊性，如图 4.2所示。这在一些先前的工作中也得到了明确的探讨 [115, 196]，因而出现了基于概率分布的边界框表示来捕捉定位模糊性。本文的定位蒸馏（LD）正是从边界框的概率分布表示的观点发展而来，这种表示最初设计用于通用目标检测，并包含丰富的定位信息。

图 4.3展示了本文的定位蒸馏框架。对于给定的目标检测器，本文按照 [115, 198]的方法将边界框表示从四元表示转换为概率分布。设 $e \in \mathcal{B}$ 为边界框的一个回归变量，其回归范围为 $[e_{\min}, e_{\max}]$。边界框分布将连续的回归范围量化为具有 n 个子区间的均匀离散变量 $\boldsymbol{e} = [e_0, e_1, \cdots, e_n] \in \mathbb{R}^{n+1}$，其中 $e_0 = e_{\min}$ 和 $e_n = e_{\max}$。定位头预测了 $n+1$ 个 logits，$\boldsymbol{z} = \{z_0, z_1, \cdots, z_n\}$，对应于子区间 $\{e_0, e_1, \cdots, e_n\}$ 的端点。通过使用 SoftMax函数，可以将给定边界框的每条边表示为概率分布。对于子区间的数量 n，本文遵循 GFocal [115]的设置，推荐选择的 n 值为 $8 \sim 16$。与 [115, 198]不同，本文使用广义 SoftMax函数 $\mathcal{S}(\cdot,\tau)$ 将 $\boldsymbol{z}_S$ 和 $\boldsymbol{z}_T$ 转换为概率分布 $\boldsymbol{p}_\tau$ 和 $\boldsymbol{q}_\tau$。注意到当 $\tau = 1$ 时，它等价于原始的 SoftMax函数。当 $\tau \to 0$ 时，它趋于狄拉克 δ 分布。当 $\tau \to \infty$ 时，它趋于均匀分布。经验上，设置 $\tau > 1$ 可以使分布变得平滑，使边界框分布携带更多信息。对于边界框

表示 $\boldsymbol{e}$ 的两个概率向量 $\boldsymbol{p}_\tau, \boldsymbol{q}_\tau \in \mathbb{R}^n$，本文通过以下损失函数进行定位蒸馏以衡量它们的相似性：

$$\mathcal{L}_{LD}^{\boldsymbol{e}} = \mathcal{H}(\boldsymbol{p}_\tau, \boldsymbol{q}_\tau) \tag{4.4}$$

$$= \mathcal{H}(\mathcal{S}(\boldsymbol{z}_S, \tau), \mathcal{S}(\boldsymbol{z}_T, \tau)). \tag{4.5}$$

然后，对于边界框 $\mathcal{B}$ 的所有四条边，定位蒸馏可以被表示为：

$$\mathcal{L}_{LD}(\mathcal{B}_S, \mathcal{B}_T) = \sum_{\boldsymbol{e} \in \mathcal{B}} \mathcal{L}_{LD}^{\boldsymbol{e}}, \tag{4.6}$$

其中，$\mathcal{B}_S, \mathcal{B}_T$ 分别是学生模型和教师模型预测的边界框。

定位蒸馏也可以灵活地用于蒸馏有向目标检测器。参数回归是经典的基于密集回归的有向目标检测中最常用的方式 [77–80]。例如常用的旋转边界框表示 $\mathcal{B} = \{\delta_x, \delta_y, \delta_w, \delta_h, \delta_\theta\}$，其中 δ_θ 表示编码的旋转角度。为了进行旋转定位蒸馏，本文同样为每个边界框表示变量生成回归范围的下限和上限 $[e_{\min}, e_{\max}]$，其中 $e \in \mathcal{B}$。需要注意的是，旋转角度预测 δ_θ 通常具有与 $\delta_x, \delta_y, \delta_w, \delta_h$ 不同的回归范围。因此，为它们设置不同的回归范围的下限和上限。实际应用时，$[e_{min}, e_{max}] \subset [-5, 5]$ 是一个可以接受的选择。随后，本文将旋转边界框转换为旋转边界框概率分布，正如前文针对水平边界框所描述的那样。最后，根据公式 4.6计算旋转边界框分布的 LD损失。

4.3 选择性区域蒸馏法

以往的知识蒸馏方法大多通过最小化 l_2 损失来使得学生网络的深层特征模仿教师网络的特征。然而，一个直接的问题出现了：是否应该毫无区别地使用整个模仿区域来蒸馏混合知识？根据本文的观察，答案是否定的。本节提出了一种有价值定位区域（Valuable Localization Region，VLR）的概念，以进一步提高蒸馏效率。本文相信这将是训练更好的学生检测器的一种有希望的方法。

具体来说，蒸馏区域被分为两部分，主蒸馏区域和有价值定位区域。主蒸馏区域由标签分配直接确定，即检测头的正样本位置。有价值定位区域则可以通过算法 1获得。首先，本文计算所有锚框 $\boldsymbol{B}^a$ 与真实边界框 $\boldsymbol{B}^{gt}$ 之间的 DIoU [201]矩阵 $\boldsymbol{X}$。然后，本文将 DIoU的下界设定为 $\alpha_{vlr} = \gamma \alpha_{pos}$，其中 α_{pos} 是标签分配的正样本 IoU阈值。那么 VLR可以定义为 $\boldsymbol{V} = \alpha_{vlr} \leqslant \boldsymbol{X} \leqslant \alpha_{pos}$。上述方法只有一个超参数 $\gamma \leqslant 1$，它控制着 VLR的范围。当 $\gamma = 0$ 时，所有预设锚框与

Algorithm 1 有价值定位区域

Require: 一组锚框 $\boldsymbol{B}^a = \{\mathcal{B}_i^a\}$ 和一组真实框 $\boldsymbol{B}^{gt} = \{\mathcal{B}_j^{gt}\}$，其中 $1 \leqslant i \leqslant I$，$1 \leqslant j \leqslant J$。标签分配的正样本阈值为 α_{pos}。

Ensure: $\boldsymbol{V} = \{v_{ij}\}_{I \times J}$，其中 $v_{ij} \in 0,1$，编码了 VLR的最终位置，其中 1 表示 VLR，0 表示忽略。

1: 计算 DIoU矩阵 $\boldsymbol{X} = \{x_{ij}\}_{I \times J}$，其中 $x_{ij} = DIoU(\mathcal{B}_i^a, \mathcal{B}_j^{gt})$.
2: $\alpha_{vlr} = \gamma \alpha_{pos}$.
3: 使用 $\boldsymbol{V} = \{\alpha_{vlr} \leqslant \boldsymbol{X} \leqslant \alpha_{pos}\}$ 筛选区域.
4: **返回** $\boldsymbol{V}$

真实框之间的 DIoU满足 $0 \leqslant x_{ij} \leqslant \alpha_{pos}$ 的位置将被确定为 VLR。而当 $\gamma \to 1$ 时，VLR将逐渐收缩为空集。本文选择使用 DIoU [201]，因为它对于靠近物体中心的位置赋予更高的优先级。

类似于标签分配，该方法在多层 FPN层级上为每个位置分配属性。通过这种方式，一些位于真实框之外的位置也会被考虑进来。因此，实际上可以将 VLR视为主蒸馏区域的向外延伸。注意到对于无锚检测器（如 FCOS），本文可以在特征图上使用预设的锚点，同时不改变其回归形式，以便保持定位学习仍然为无锚类型。而对于通常在每个位置设置多个锚框的基于锚框的检测器，本文会展开这些锚框以计算 DIoU矩阵，并为它们分配属性。

最后根据以上描述，用于训练学生网络 $\boldsymbol{S}$ 的 logit模仿总损失可以表示为：

$$\begin{aligned}\mathcal{L} =& \lambda_0 \mathcal{L}_{\text{cls}}(\mathcal{C}_S, \mathcal{C}^{gt}) + \lambda_1 \mathcal{L}_{\text{reg}}(\mathcal{B}_S, \mathcal{B}^{gt}) + \lambda_2 \mathcal{L}_{\text{DFL}}(\mathcal{B}_S, \mathcal{B}^{gt}) \\ &+ \lambda_3 \mathbb{I}_{\text{Main}} \mathcal{L}_{\text{LD}}(\mathcal{B}_S, \mathcal{B}_T) + \lambda_4 \mathbb{I}_{\text{VLR}} \mathcal{L}_{\text{LD}}(\mathcal{B}_S, \mathcal{B}_T) \\ &+ \lambda_5 \mathbb{I}_{\text{Main}} \mathcal{L}_{\text{KD}}(\mathcal{C}_S, \mathcal{C}_T) + \lambda_6 \mathbb{I}_{\text{VLR}} \mathcal{L}_{\text{KD}}(\mathcal{C}_S, \mathcal{C}_T),\end{aligned} \tag{4.7}$$

其中前三项与基于回归的检测器的分类和边界框回归分支完全相同，即 $\mathcal{L}_{\text{cls}}$ 是分类损失，$\mathcal{L}_{\text{reg}}$ 是边界框回归损失，$\mathcal{L}_{\text{DFL}}$ 是分布聚焦损失 [115]。 $\mathbb{I}_{\text{Main}}$ 和 $\mathbb{I}_{\text{VLR}}$ 分别是主要蒸馏区域和有价值的定位区域的蒸馏掩码。 $\mathcal{L}_{\text{KD}}$ 是 KD损失 [205]，$\mathcal{C}_S$ 和 $\mathcal{C}_T$ 分别表示学生网络和教师网络的分类头输出逻辑值，$\mathcal{C}^{gt}$ 是真实类别标签。

所有蒸馏损失将根据它们的类型被赋予相同的权重因子，即 LD损失的权重因子与边界框回归损失的权重因子相同，分类 KD损失的权重因子与分类损失的权重因子相同。值得一提的是，由于 LD损失具有足够的监督能力，DFL损失项可以被舍弃。另外，本文可以选择启用或禁用这四种类型的蒸馏损失（通过设置 $\lambda_3, \lambda_4, \lambda_5, \lambda_6$ 是否为0），以便有选择地在不同区域对学生进行蒸馏。

4.4 实验

本节将进行全面的消融实验和分析，以验证所提出的 LD和蒸馏方案在具有挑战性的大规模目标检测数据集 MS COCO [37]、经典自然场景数据集 PASCAL VOC [36]以及遥感影像数据集 DOTA [29]上的优越性。

4.4.1 实验设置

对于 COCO，本文使用 train2017（118K张图像）进行训练，val2017（5K张图像）用于验证。本文还通过提交推理结果到 COCO官方服务器，在 MS COCO test-dev 2019数据集（20K张图像）上进行了评估。所有实验是在 mmDetection [265]框架下进行的，以确保公平的实验环境。除非另有说明，本文使用 ResNet [108]作为骨干网络，并结合 FPN [105]作为颈部网络，使用 FCOS [101]风格的无锚头用于分类和定位。消融实验的训练周期设置为单尺度1× 模式，即12个 epochs。对于其他训练和测试超参数，本文完全遵循 GFocal [115]的设定，包括分类任务使用 QFL损失，边界框回归任务使用 GIoU损失等。本文使用标准的 COCO评估方式，即平均精度（AP）。所有的基准模型都采用相同的设置重新训练，以便与本文的 LD进行公平比较。

本节还提供了在另一个流行的目标检测基准测试上的实验结果，即 PASCAL VOC [36]。本文使用 VOC 07+12训练协议，即将 VOC 2007的 trainval集和 VOC 2012的 trainval集（16551张图像）联合起来进行训练，然后使用 VOC 2007的测试集（4952张图像）进行评估。初始学习率为0.01，总训练 epochs设置为4。在第3个 epoch之后，学习率会减小10倍。为了全面评估定位性能，本文报告了平均精度（AP）以及5个不同 IoU阈值下的 mAP，即 AP50、AP60、AP70、AP80 和 AP_{90}。

对于旋转 LD的评估，本文在经典的遥感图像数据集 DOTA [29]上报告了检测结果。本文遵循标准的 mmRotate [280]的训练和测试协议。训练集和验证集分别包含1403张和468张图像，在文献中这些图像是随机选择的。这些巨大的图像被裁剪成形状为 600×600 的小子图像，这与官方实现中的裁剪协议保持一致。在实践中，本文获得了大约15700个训练 patches和5300个验证 patches。除非另有说明，所有超参数都遵循 mmRotate的默认设置，以进行公平比较。本文以 AP和5个不同 IoU阈值下的 mAP为指标进行结果报告，这与 PASCAL VOC保持一致。由于内存限制，教师网络使用 ResNet-34-FPN，并进行2× 的训练周期

表 4.1: **LD中的温度参数 τ**：使用较大的 τ 值的广义 Softmax函数带来了显著的收益。本文默认将 τ 设置为10。教师网络为 ResNet-101，学生网络为 ResNet-50。

τ	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
–	40.1	58.2	43.1	23.3	44.4	52.5
1	40.3	58.2	43.4	22.4	44.0	52.4
5	40.9	58.2	44.3	23.2	**45.0**	53.2
10	**41.1**	**58.7**	**44.9**	**23.8**	44.9	**53.6**
15	40.7	58.5	44.2	23.5	44.3	53.3
20	40.5	58.3	43.7	**23.8**	44.1	53.5

（24个 epochs），而学生网络使用 ResNet-18-FPN，并进行1× 的训练周期（12个 epochs）。

4.4.2 消融实验

（1）LD的温度参数 τ

本文的 LD引入了一个超参数，即温度 τ，用于控制边界框概率分布的软化程度。表4.1报告了使用不同温度的 LD的结果，其中教师模型是具有 AP 44.7的 ResNet-101，学生模型是 ResNet-50。在这里，只采用了主蒸馏区域进行实验。与表 4.1中的第一行相比，不同的温度一致地导致更好的结果。在本文中，LD中的温度将默认设为 $\tau = 10$，并在所有其它实验中固定使用该值。

（2）LD *vs.* 伪边界框回归

教师边界框回归（TBR）损失 [207]是增强学生网络的定位头的初步尝试，即图 4.1中的伪边界框回归。 TBR损失可以表示为：

$$\mathcal{L}_{\text{TBR}} = \lambda \mathcal{L}_{\text{reg}}(\mathcal{B}^s, \mathcal{B}^{gt}), \text{if } \ell_2(\mathcal{B}^s, \mathcal{B}^{gt}) + \varepsilon > \ell_2(\mathcal{B}^t, \mathcal{B}^{gt}), \tag{4.8}$$

其中，$\mathcal{B}^s$ 和 $\mathcal{B}^t$ 分别表示学生和教师的预测边界框， $\mathcal{B}^{gt}$ 表示真实边界框， ε 是预定义的边界， $\mathcal{L}_{\text{reg}}$ 表示 GIoU损失 [200]。在这里，只采用了主蒸馏区域进行实验。从表 4.2中，可以看到当在公式 4.8中使用适当的阈值 $\varepsilon = 0.1$ 时，TBR损失确实产生了性能增益（+0.4 AP和+0.7 AP_{75}）。然而，TBR损失使用了粗糙的边界框表示，其中不包含检测器的任何定位不确定性信息，从而导致次优的结果。相反，本文的 LD直接获得了41.1的 AP和44.9的 AP_{75}，因为它利用了包含丰富定位知识的边界框概率分布。

表 4.2: **LD *vs*. 伪边界框回归 [207]**：相比于伪边界框回归，LD能够更有效地传递定位知识。教师网络为 ResNet-101，学生网络为 ResNet-50。

ε	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
–	40.1	58.2	43.1	23.3	44.4	52.5
0.1	40.5	58.3	43.8	23.0	44.2	52.7
0.2	40.2	58.2	43.6	23.1	44.0	53.0
0.3	40.1	58.4	43.1	23.6	43.9	52.5
0.4	40.3	58.4	43.4	22.8	44.0	52.6
LD	**41.1**	**58.7**	**44.9**	**23.8**	**44.9**	**53.6**

（3）单独 LD

为进一步研究 LD的性能，在该实验中，本文不使用真实边界框来训练学生检测器，即禁用方程公式 4.7中的边界框回归损失 $\mathcal{L}_{\text{reg}}$ 和分布焦点损失 $\mathcal{L}_{\text{DFL}}$，以及设 $\lambda_4, \lambda_5, \lambda_6 = 0$，从而得到单独 LD。从表4.3中可以看到，当仅使用主蒸馏区域内的 LD来训练学生 ResNet-18时，获得了36.4 AP和39.3 AP_{75}。令人惊讶的是，单独 LD的结果要高于使用 $\mathcal{L}_{\text{reg}}$ 和 $\mathcal{L}_{\text{DFL}}$（第一行）的结果，这说明了仅仅依靠教师模型的定位结果，LD就足以帮助学生检测器获得出色的定位能力。而在同时添加 $\mathcal{L}_{\text{reg}}$ 和 $\mathcal{L}_{\text{DFL}}$ 的情况下，AP仅增加了0.1，这表明教师检测器学习到的概率分布比真实框具有更好的定位监督效果。

表 4.3: **单独 LD的结果**：可产生比真实框更强力的定位监督。

	教师	学生	$\mathcal{L}_{\text{LD}}$	$\mathcal{L}_{\text{reg}}$ and $\mathcal{L}_{\text{DFL}}$	AP	AP_{75}
	R-101	**R**-18		✓	35.8	38.2
			✓		36.4	39.3
			✓	✓	36.5	39.3

（4）自我蒸馏

在知识蒸馏的一般设定下，学生模型 $\boldsymbol{S}$ 一般比老师模型 $\boldsymbol{T}$ 更轻量，从而学习到紧致高效的模型。然而若想要提高最大模型的性能，这一设定就会受到应用限制。最近，人们不断观察到自我蒸馏在分类 [281, 282]中具有积极作用。对于目标检测，同样令人欣慰的是，$\boldsymbol{S} = \boldsymbol{T}$ 的 Self-LD也能带来性能提升。而为何自我蒸馏能够提高模型精度，Mobahi H.等人 [283]通过对希尔伯特空间中的自我蒸馏进行理论分析揭开了其中的神秘面纱。已被证明的是，几轮的自我蒸馏可

以减少过拟合，因为它会引入正则化。然而，持续的自我蒸馏可能会导致欠拟合。因此本文只需进行一次自我 LD。如表 4.4所示，主蒸馏区域上的 Self-LD将性能提升了+0.3 AP、+0.3 AP_{75}（对于 ResNet-18）、+0.5 AP、+0.7 AP_{75}（对于 ResNet-50）以及 +0.6 AP、+0.7 AP_{75}（对于 ResNeXt-101-32x4d-DCN）。自我蒸馏展示了本文 LD的普遍性，即当教师模型与学生模型具有相同的规模时，定位知识仍然可以被迁移。

表 4.4: **自我蒸馏：**当学生模型与教师模型的模型规模相同时，LD依然可以提高检测性能。仅使用了主蒸馏区域。

检测器	自我蒸馏	AP	AP_{75}
ResNet-18		35.8	38.2
	✓	36.1 (+0.3)	38.5
ResNet-50		40.1	43.1
	✓	40.6 (+0.5)	43.8
ResNeXt-101-32x4d-DCN		46.9	51.1
	✓	47.5 (+0.6)	51.8

（5）VLR中的不同 γ

新引入的 VLR具有参数 γ，它控制 VLR的范围。如表 4.5所示，当 γ 的取值范围从0到0.5时，AP保持稳定。在这个范围内，AP的变化大约在0.1左右。随着 γ 的增加，VLR逐渐收缩为空。性能也逐渐下降到41.1，即仅在主蒸馏区域上进行 LD。对参数 γ 的敏感性分析实验证明，在 VLR上进行 LD对性能有积极的影响。在其余的实验中，为了简单起见，本文将 γ 设置为0.25。

表 4.5: **γ 在 VLR中的作用：**在有价值定位区域上进行 LD对性能有积极的影响。本文默认将 γ 设置为0.25。教师网络为 ResNet-101，学生网络为 ResNet-50。

γ	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
–	40.1	58.2	43.1	23.3	44.4	52.5
1	41.1	58.7	44.9	23.8	44.9	53.6
0.75	41.2	58.8	44.9	23.6	45.4	53.5
0.5	41.7	59.4	45.3	24.2	45.6	54.2
0.25	**41.8**	**59.5**	**45.4**	24.2	45.8	**54.9**
0	41.7	**59.5**	**45.4**	**24.5**	**45.9**	54.0

（6）选择性区域蒸馏法

关于分类 KD和 LD的作用以及它们的优选区域，有几个有趣的观察结果。表 4.6中报告了相关的消融实验结果，其中“Main”表示在主要蒸馏区域上进行 logit模仿，即标签分配的正样本位置，“VLR”表示有价值定位区域。对于 MS COCO数据集，可以看到进行“Main LD”、“VLR LD”和“Main KD”都有助于学生网络的性能提升。这表明主要蒸馏区域包含了有价值的分类和定位知识，而分类 KD相比 LD的效果较差。然后，本文将分类 KD扩展到更大的范围，即 VLR。然而，实验观察到进一步将“VLR KD”引入并没有带来任何改进（表 4.6的最后两行）。这就是为什么本文采用选择性区域蒸馏法来处理 COCO数据集的主要原因。

表 4.6: **分类 KD和 LD的选择性区域蒸馏评估**：COCO数据集的师生对使用 ResNet-101→ResNet-50，VOC 07+12数据集的师生对是 ResNet-101→ResNet-18。

LD		KD		MS COCO val2017			VOC 07+12		
Main	VLR	Main	VLR	AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}
				40.1	58.2	43.1	51.8	75.8	56.3
✓				41.1	58.7	44.9	53.0	75.9	57.6
✓	✓			41.8	59.5	45.4	53.4	76.3	**58.3**
✓	✓	✓		**42.1**	**60.3**	**45.6**	53.1	76.8	57.6
✓	✓	✓	✓	42.0	60.0	45.4	**53.7**	**77.3**	58.2

接下来，本文进一步报告了在 PASCAL VOC数据集上的 KD和 LD的作用。从表 4.6中可以看出，将定位知识转移给主要蒸馏区域和 VLR都是有益的。然而，由于不同的知识分布模式，分类知识蒸馏也显示出类似的性能下降。通过比较表 4.6中的第3行和第4行，“Main KD”导致了性能下降，而“VLR KD”对学生网络产生了积极影响。这表明选择性区域蒸馏可以在它们各自有利的区域充分发挥 KD和 LD的优势。

（7）轻量级检测器的结果

表 4.7报告了本文的蒸馏方案 (COCO上的“Main LD + VLR LD + Main KD”)，其中对一系列轻量级学生网络进行了蒸馏，包括 ResNet-18、ResNet-34和 ResNet-50。对于所有给定的学生网络，本文的 LD都可以稳定地提高检测性能，而无需任何复杂的操作。从这些结果中，可以看到本文的 LD分别将

ResNet-18、ResNet-34和 ResNet-50的 AP提高了+1.7、+2.1和+2.0，将 AP_{75} 提高了+2.2、+2.4和+2.5。

表 4.7: **轻量级检测器的 LD的定量结果：**教师模型为 ResNet-101。结果报告在 MS COCO val2017数据集上。

学生	LD	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
ResNet-18		35.8	53.1	38.2	18.9	38.9	47.9
	✓	37.5	54.7	40.4	20.2	41.2	49.4
ResNet-34		38.9	56.6	42.2	21.5	42.8	51.4
	✓	41.0	58.6	44.6	23.2	45.0	54.2
ResNet-50		40.1	58.2	43.1	23.3	44.4	52.5
	✓	42.1	60.3	45.6	24.5	46.2	54.8

（8）其它密集检测器的结果

本文的 LD可以灵活地应用于其他密集目标检测器，包括基于锚框和不基于锚框的类型。本文采用分而治之的蒸馏方案将 LD应用于几种最近流行的检测器，如 RetinaNet [106]（基于锚框）、FCOS [101]（不基于锚框）和 ATSS [114]（基于锚框）。根据表 4.8中的结果，可以看到本文的 LD可以稳定地将基线模型的 AP提高约2分。

表 4.8: **LD在各种流行的密集目标检测器上的定量结果：**教师模型是 ResNet-101，学生模型是 ResNet-50。结果报告在 MS COCO val2017数据集上。

学生	LD	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
RetinaNet [106]		36.9	54.3	39.8	21.2	40.8	48.4
	✓	39.0	56.4	42.4	23.1	43.2	51.1
FCOS [101]		38.6	57.2	41.5	22.4	42.2	49.8
	✓	40.6	58.4	44.1	24.3	44.1	52.3
ATSS [114]		39.2	57.3	42.4	22.7	43.1	51.5
	✓	41.6	59.3	45.3	25.2	45.2	53.3

（9）LD对模型大小、计算量和推理速度的影响

由于 LD需要将边界框表示从狄拉克 δ 分布转换为概率分布，因此对检测器的唯一修改在于定位头输出通道，从 $H \times W \times 4$ 转换到 $H \times W \times 4n$。 表 4.9中研究了这种对模型大小、计算量（FLOPs）和推理速度（FPS）的影响。可以看

到，对于 FCOS和 ATSS这两种目标检测器，LD可以显着提高性能，而模型大小和计算量的增加可以忽略不计。而对于 RetinaNet，LD会带来 FPS的轻微下降。这主要是因为 RetinaNet为每个特征点预测9个框，这导致 RetinaNet的定位头输出通道修改将从 $H \times W \times (9 \times 4)$ 到 $H \times W \times (9 \times 4n)$，是 FCOS和 ATSS的9倍。至于 GFocal，由于它本身即是采用概率分布表示的边界框建模，因而 LD可获得免费的性能提升。

表 4.9: **LD对模型大小、计算量和推理速度的影响**：FPS是使用 RTX 2080 Ti GPU测量的，并在3次运行中取平均值。

	LD	图像尺寸	参数量	FLOPs	FPS	AP
RetinaNet		1333×800	37.74M	239.32G	21.0	36.9
	✓	1333×800	39.07M	267.64G	19.6	39.0
FCOS		1333×800	32.02M	200.50G	22.3	38.6
	✓	1333×800	32.17M	203.60G	22.4	40.6
ATSS		1333×800	32.07M	205.21G	21.9	39.2
	✓	1333×800	32.22M	208.36G	21.9	41.6
GFocal		1333×800	32.22M	208.31G	21.9	40.1
	✓	1333×800	32.22M	208.31G	21.9	42.1

（10）有向目标检测器的结果

作为 LD的直接扩展，旋转边界框需要学习额外的旋转角度概率分布。本文对两个有向目标检测器进行了必要的最小修改：1）基于密集回归的旋转检测器 Rotated-RetinaNet [106]，这是许多有向目标检测器的基础；2）最近流行的2D高斯分布建模检测器 GWD [79]。本实验遵循 mmRotate [280]的训练和测试协议。这里使用 ResNet-34作为教师模型，使用 ResNet-18作为学生模型以节省 GPU内存。结果报告在 DOTA-v1.0 [29]验证集上。如表 4.10所展示，旋转 LD也能够成功应用于有向目标检测器，并在遥感图像目标检测任务中取得了显著的性能提升。特别是，在更严格的 IoU阈值下，如 AP70、AP80、AP_{90}，本文方法获得了令人印象深刻的提升。这显示了本文的定位蒸馏的卓越兼容性，它不仅可以应用于水平边界框，还可以应用于旋转边界框。此外，值得一提的是，本文的定位蒸馏不依赖于边界框的表示方式以及建模的优化方式，即对于水平边界框预测使用的 IoU-based loss [200, 201]以及旋转边界框预测使用的2D高斯建模 [79]兼容有效。

表 4.10: **旋转 LD在流行的有向目标检测器上的定量结果：**教师模型是 ResNet-34，学生模型是 ResNet-18。结果报告在 DOTA-v1.0验证集上。

学生	AP	AP_{50}	AP_{60}	AP_{70}	AP_{80}	AP_{90}
R-RetinaNet [106]	33.7	58.0	54.5	42.3	22.9	4.7
LD (ours)	39.1	63.8	61.1	48.8	28.7	8.8
GWD [79]	37.1	63.1	60.1	46.7	24.7	6.2
LD (ours)	40.2	66.4	63.6	50.3	28.2	8.5

（11）LD在 PASCAL VOC上的结果

本文进一步在 PASCAL VOC [36]上进行实验，以验证 LD的普适性。这里仅采用主蒸馏区域。 表 4.11展示了 LD可持续提升学生 ResNet-18的性能。值得注意的是，在高 IoU阈值下的严格指标，LD的表现仍明显优于基线，例如 AP_{80} 和 AP_{90}。

表 4.11: **LD在 PASCAL VOC [36]上的结果：**结果报告在 VOC 2007测试集上。

教师	学生	AP	AP_{50}	AP_{60}	AP_{70}	AP_{80}	AP_{90}
–	ResNet-18	51.8	75.8	72.0	62.9	46.5	20.6
ResNet-34		52.9	75.7	72.2	**64.5**	**49.1**	22.0
ResNet-50		52.8	75.4	72.1	64.0	49.0	**23.0**
ResNet-101		**53.0**	**75.9**	**72.4**	64.0	48.9	22.7
ResNet-101-DCN		52.8	75.3	72.0	64.2	49.0	22.1

（12）与其它先进目标检测器的比较

本节继续展示 LD与最先进的密集目标检测器的性能比较，通过在 GFocalV2 [129]上应用本文的 LD来进一步提升性能。对于 COCO val2017数据集，由于大多数先前的方法都使用 ResNet-50-FPN骨干网络，采用单尺度 1× 的训练周期（12个 epochs）进行验证，本文也在这个设置下报告结果，以便进行公平比较。对于 COCO test-dev 2019数据集，本文按照先前的工作 [129]，开启多尺度训练，分辨率为 $1333 \times [480 : 960]$，并使用 2× 的训练周期（24个 epochs）。训练在一台含8块 GPU的设备上进行，每块 GPU的批量大小为2，初始学习率为0.01，以进行公平比较。在推理阶段，采用单尺度测试（$[1333 \times 800]$分辨率）。对于不同的学生网络，如 ResNet-50、ResNet-101和 ResNeXt-101-32x4d-DCN [267,284]，本文分别选择 ResNet-101、ResNet-101-DCN和 Res2Net-101-DCN [285]作为教师

模型，以确保教师模型的性能表现总是优于学生模型。

表 4.12: **在 COCO val2017和 test-dev2019上与先进目标检测器的比较**：“1×”：单尺度训练12epochs。“2×”：多尺度训练24epochs。**TS**: 训练周期。

Method	**TS**	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
ResNet-50 backbone on val2017							
RetinaNet [106]	1×	36.9	54.3	39.8	21.2	40.8	48.4
FSAF [286]	1×	37.2	57.2	39.4	21.0	41.2	49.7
FCOS [101]	1×	38.6	57.2	41.5	22.4	42.2	49.8
SAPD [287]	1×	38.8	58.7	41.3	22.5	42.6	50.8
SABL [288]	1×	38.8	57.9	41.6	22.7	42.9	50.5
ATSS [114]	1×	39.2	57.3	42.4	22.7	43.1	51.5
BorderDet [289]	1×	41.4	59.4	44.5	23.6	45.1	54.6
AutoAssign [290]	1×	40.5	59.8	43.9	23.1	44.7	52.9
PAA [291]	1×	40.4	58.4	43.9	22.9	44.3	54.0
OTA [138]	1×	40.7	58.4	44.3	23.2	45.0	53.6
GFocal [115]	1×	40.1	58.2	43.1	23.3	44.4	52.5
GFocalV2 [129]	1×	41.1	58.8	44.9	23.5	44.9	53.3
LD (ours)	1×	**42.7**	**60.2**	**46.7**	**25.0**	**46.4**	**55.1**
ResNet-101 backbone on test-dev 2019							
RetinaNet [106]	2×	39.1	59.1	42.3	21.8	42.7	50.2
FSAF [286]	2×	40.9	61.5	44.0	24.0	44.2	51.3
FCOS [101]	2×	41.5	60.7	45.0	24.4	44.8	51.6
SAPD [287]	2×	43.5	63.6	46.5	24.9	46.8	54.6
SABL [288]	2×	43.6	62.9	47.2	27.4	48.4	55.7
ATSS [114]	2×	43.6	62.1	47.4	26.1	47.0	53.6
BorderDet [289]	2×	45.4	64.1	48.8	26.7	48.3	56.5
AutoAssign [290]	2×	44.5	64.3	48.4	25.9	47.4	55.0
PAA [291]	2×	44.8	63.3	48.7	26.5	48.8	56.3
OTA [138]	2×	45.3	63.5	49.3	26.9	48.8	56.1
GFocal [115]	2×	45.0	63.7	48.9	27.2	48.8	54.5
GFocalV2 [129]	2×	46.0	64.1	50.2	27.6	49.6	56.5
LD (ours)	2×	**47.1**	**65.0**	**51.4**	**28.3**	**50.9**	**58.5**
ResNeXt-101-32x4d-DCN backbone on test-dev 2019							
SAPD [287]	2×	46.6	66.6	50.0	27.3	49.7	60.7
GFocal [115]	2×	48.2	67.4	52.6	29.2	51.7	60.2
GFocalV2 [129]	2×	49.0	67.6	53.4	29.8	52.3	61.8
LD (ours)	2×	**50.5**	**69.0**	**55.3**	**30.9**	**54.4**	**63.4**

如表 4.12所示，当使用 ResNet-50-FPN骨干网络时，本文的 LD将最先进的 GFocalV2的 AP分数提高了+1.6，AP_{75} 分数提高了+1.8。当使用 ResNet-101-FPN和 ResNeXt-101-32x4d-DCN并以多尺度 2× 训练时，本文方法实现了最高的AP分数，分别为47.1和50.5。这一结果在相同的主干网络、颈部网络和测试设置下，优于所有现有的密集目标检测器。更重要的是，LD不会引入任何额外的网络参数和计算开销，因此可以保证与 GFocalV2完全相同的推理速度。

（13）LD的检测效果可视化

图 4.4展示了几个使用 GFocalv2和本文的 LD的检测效果示例，其中第三列与第四列结果由 IoU阈值为0.95的 NMS获得。可以看到原始 GFocalv2的检测框具有不同的定位质量，并且在 NMS后仍可能存在冗余框。相比之下，本文的 LD有助于提高检测框的定位质量，这使得冗余框能够更容易被 NMS算法抑制。

4.5 定位蒸馏的理论性质

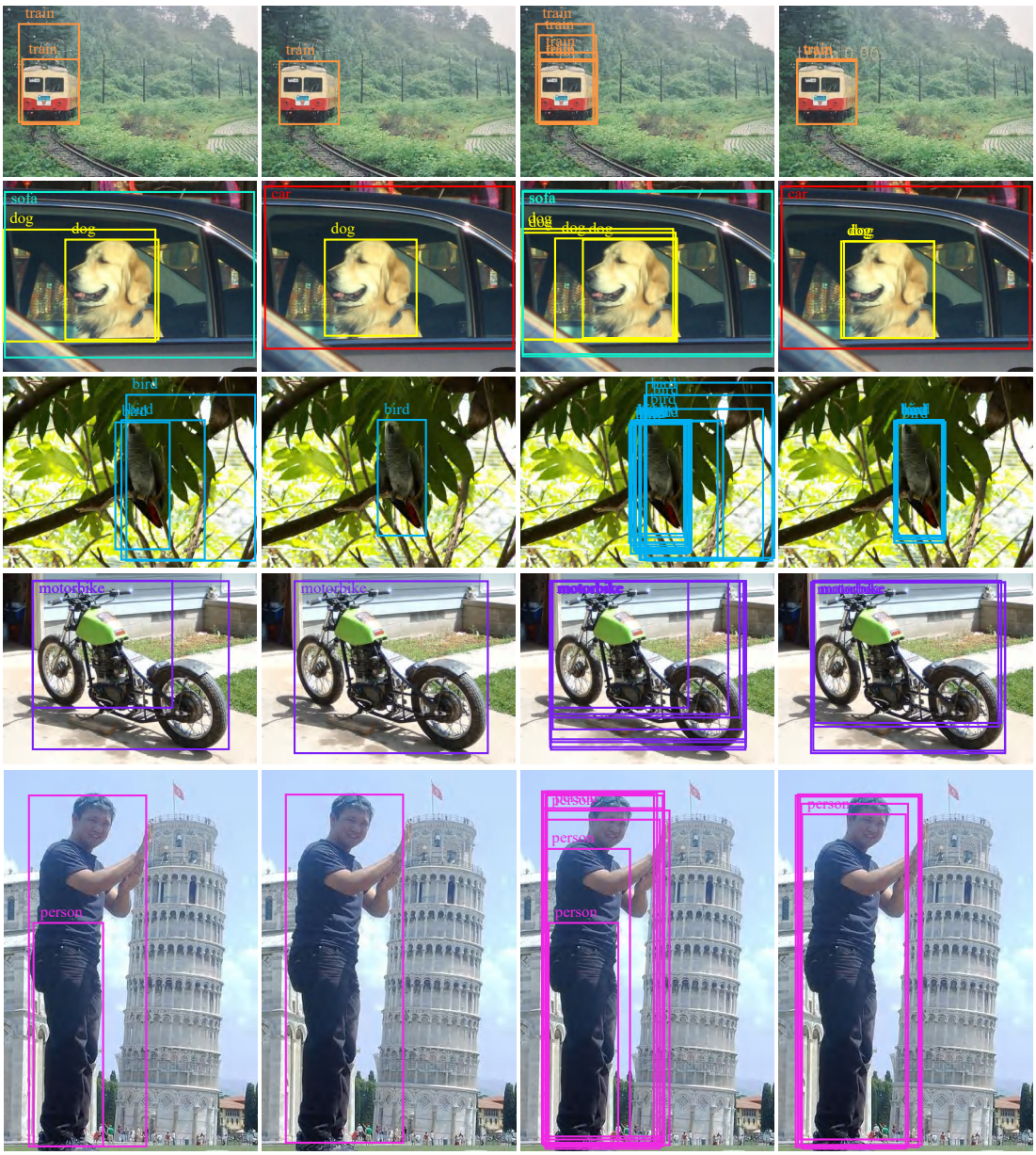
经过以上对定位蒸馏的充分消融实验，本文展示了定位蒸馏（LD）在目标检测知识蒸馏中起到重要作用，并且分类知识与定位知识可因地制宜地传递给学生。然而 LD的工作机理，以及其与传统的分类 KD的联系仍不清楚。本节的内容将 LD的理论性质进行探讨。

4.5.1 LD的性质

实际上，由公式 4.4可以看到，LD保持了标准 logit模仿的形式。一个可能会问的问题是：LD是否也继承了分类 KD的特性，特别是优化过程中的特性？与分类任务中唯一的整数被视为真实标签不同，定位任务的真实标签是一个浮点数 e^*，其值在某个小区间 $[e_i, e_{i+1}]$ 内。这意味着在定位蒸馏过程中，我们需要处理连续值的回归目标，而不是离散的类别标签。在接下来，本文展示了 LD的一个重要特性，证明它可以继承分类 KD所具有的优化效果。

命题 1 *设 $\boldsymbol{s}$ 为学生的预测概率向量，u_1 和 u_2 是两个常数，满足 $u_1+u_2=1$. 我们有:*

1. *若 $\boldsymbol{p}$ 和 $\boldsymbol{q}$ 是两个分类概率向量，LD对线性组合 $\boldsymbol{l}=u_1\boldsymbol{p}+u_2\boldsymbol{q}$ 的效应等于 KD对 $\boldsymbol{p}$ 和 $\boldsymbol{q}$ 效果的线性组合;*
2. *若 $\boldsymbol{l}$ 是一个定位概率向量， LD对 $\boldsymbol{l}$ 的效果等于作用在其分解 $\boldsymbol{p}$ 和 $\boldsymbol{q}$ 上的*



GFocalv2 LD GFocalv2 “0.95” LD “0.95”

图 4.4: GFocalv2和 LD的检测结果比较。“0.95” 表示 GFocalv2中的 NMS的IoU阈值由0.6增加到0.95。最好在彩色图中查看。

两个 *KD*效果的线性组合。

上述两个观点具有相同的表达式，

$$\partial LD_i^{\boldsymbol{l}} = u_1 \partial KD_i^{\boldsymbol{p}} + u_2 \partial KD_i^{\boldsymbol{q}}, \tag{4.9}$$

其中 $\partial KD_i^{\boldsymbol{p}}$ 表示针对给定的逻辑值 z_i，两个概率 $\boldsymbol{s}, \boldsymbol{p}$ 的 *KD*损失的导数， $\partial LD_i^{\boldsymbol{p}}$

同样表示针对给定的逻辑值 z_i，概率 $\boldsymbol{p}$ 的 *LD*损失的导数。

上述命题提供了 LD与分类 KD之间的理论联系。它表明，对于一个浮点数定位问题，LD的优化效果在功能上等价于作用在整数位置分类问题上的两个KD效果。

在证明上述命题之前，本文首先给出一些符号：$\boldsymbol{g}^i=[g_1,g_2,\cdots,g_n]$ 是一个分类标签向量，其中只在第 i 个位置上 $g_i=1$ 而在其它位置上为 0.

$\boldsymbol{e}=[e_1,e_2,\cdots,e_n]\in\mathbb{R}^n$ 是一个均匀分割的离散向量，用于表示回归范围 $[e_{\min},e_{\max}]$ 被分割后的小区间端点值。

交叉熵损失（Cross Entropy loss，CE损失）关于其中某一个 logit $z_i\in\boldsymbol{z}_S$, $i\in\{1,2,\cdots,n\}$ 的梯度为:

$$\frac{\partial\mathcal{L}_{\mathrm{CE}}}{\partial z_i}=p_i-g_i, \tag{4.10}$$

其中 p_i 是位置 i 处的预测分类概率，z_S 是学生输出 logit向量.

加上 CE损失后，KD损失关于其中某一个 logit $z_i\in\boldsymbol{z}_S$ 的梯度为:

$$\frac{\partial\mathcal{L}^{KD}}{\partial z_i}=\gamma(p_i-g_i)+\frac{\lambda}{\tau}(p_{\tau i}-q_{\tau i}), \tag{4.11}$$

其中 γ 与 λ 是 CE损失与 KD损失权重，τ 温度参数。

本文沿用了 [292]中的记号，记 $\frac{\partial\mathcal{L}_{\mathrm{CE}}}{\partial z_i}$ 为 ∂_i，以及记 $\frac{\partial\mathcal{L}^{KD}}{\partial z_i}$ 为 $\frac{\partial_i^{KD}}{\partial_i}$.

那么公式 4.11与公式 4.10的比值表明了 KD是在 logit空间内对 CE损失的梯度重缩放。

定义 1 设 $\boldsymbol{p}\in\mathbb{R}^n$ 是一个网络模型的预测概率向量，$M_i>0,i\in\{1,2,\cdots,n\}$ 一些预定义阈值，所执行的任务为 T. 如果 $\boldsymbol{p}$ 到其对应的真实值向量 $\boldsymbol{g}^i$ 的距离是 M_i 有界的，则称 $\boldsymbol{p}$ 对任务 **T** 是 M_i 表现良好的。

引理 1 如果两个预测概率向量 $\boldsymbol{p},\boldsymbol{q}$ 分别对整数位置分类任务是 M_i 表现良好的与 M_j 表现良好的，$\boldsymbol{g}^i,\boldsymbol{g}^j$ 是它们所对应的真实值向量，则它们的线性组合 $u_1\boldsymbol{p}+u_2\boldsymbol{q}$ 是对浮点数位置定位任务 M 表现良好的，其对应的真实值是 $y=u_1e_i+u_2e_j$. 这里 $M=\max\{M_i,M_j\}$，$u_1+u_2=1$.

证明. 根据定义1，两个距离满足 $d(\boldsymbol{p},\boldsymbol{g}^i)\leqslant M_i, d(\boldsymbol{q},\boldsymbol{g}^j)\leqslant M_j$，其中 $\boldsymbol{g}^i\neq\boldsymbol{g}^j$. 注意到 $d(\cdot,\cdot)$ 可以是任意距离度量，如 l_2 距离。

一个浮点数位置定位任务需要一个概率，其可以通过线性插值 $\boldsymbol{l}=u_1\boldsymbol{p}+u_2\boldsymbol{q}$ 得到，其真实值向量为 $\boldsymbol{g}=u_1\boldsymbol{g}^i+u_2\boldsymbol{g}^j$.

于是，我们有

$$
\begin{aligned}
d(\boldsymbol{l},\boldsymbol{g}) &= d(u_1\boldsymbol{p}+u_2\boldsymbol{q}, u_1\boldsymbol{g}^i+u_2\boldsymbol{g}^j) && (4.12)\\
&\leqslant d(u_1\boldsymbol{p}+u_2\boldsymbol{q}, u_1\boldsymbol{g}^i+u_2\boldsymbol{q}) \\
&\quad + d(u_1\boldsymbol{g}^i+u_2\boldsymbol{q}, u_1\boldsymbol{g}^i+u_2\boldsymbol{g}^j) && (4.13)\\
&= d(u_1\boldsymbol{p}, u_1\boldsymbol{g}^i) + d(u_2\boldsymbol{q}, u_2\boldsymbol{g}^j) && (4.14)\\
&\leqslant u_1 M_i + u_2 M_j && (4.15)\\
&\leqslant \max\{M_i, M_j\} && (4.16)\\
&= M. && (4.17)
\end{aligned}
$$

因此，该网络在浮点数位置定位任务上为 M 表现良好的。

□

引理 2 如果 $\boldsymbol{l}$ 是一个定位概率向量，有着真实值 $y = u_1 e_i + u_2 e_j$，其中 $u_1 + u_2 = 1$，则 $\boldsymbol{l}$ 可以被分解为两个分类概率向量 $\boldsymbol{p}$ 与 $\boldsymbol{q}$，分别有着真实值向量 $\boldsymbol{g}^i$ 与 $\boldsymbol{g}^j$.

证明. 设 $\boldsymbol{l} \in \mathbb{R}^n$ 是一个预测定位概率向量，$\boldsymbol{g}$ 是其对应的真实值向量。

由于 $\boldsymbol{l}$ 是浮点数，因而很容易将 $\boldsymbol{g}$ 分解出两个整数位置真实值向量 $\boldsymbol{g}^i$ 与 $\boldsymbol{g}^j$，满足 $\boldsymbol{g} = u_1\boldsymbol{g}^i + u_2\boldsymbol{g}^j$.

$\boldsymbol{l}$ 的分解的存在性：

为了将 $\boldsymbol{l}$ 分解为两个分类概率向量 $\boldsymbol{p}$ 与 $\boldsymbol{q}$，且满足 $\boldsymbol{l} = u_1\boldsymbol{p} + u_2\boldsymbol{q}$，我们需要解以下线性方程组，并只需证明该方程组有解：

$$
AX = \boldsymbol{b} \Longleftrightarrow \begin{cases}
\sum\limits_i p_i = 1, \\
\sum\limits_i q_i = 1, \\
u_1 p_1 + u_2 q_1 = l_1, \\
u_1 p_2 + u_2 q_2 = l_2, \\
\vdots \\
u_1 p_n + u_2 q_n = l_n,
\end{cases} \tag{4.18}
$$

其中 $X=(p_1,p_2,\cdots,p_n,q_1,q_2,\cdots,q_n)^T$，增广矩阵 $(A,\boldsymbol{b})$ 为：

$$\begin{pmatrix} 1 & 1 & 1 & \cdots & 1 & 0 & 0 & 0 & \cdots & 0 & 1 \\ 0 & 0 & 0 & \cdots & 0 & 1 & 1 & 1 & \cdots & 1 & 1 \\ u_1 & 0 & 0 & \cdots & 0 & u_2 & 0 & 0 & \cdots & 0 & l_1 \\ 0 & u_1 & 0 & \cdots & 0 & 0 & u_2 & 0 & \cdots & 0 & l_2 \\ 0 & 0 & u_1 & \cdots & 0 & 0 & 0 & u_2 & \cdots & 0 & l_3 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & u_1 & 0 & 0 & 0 & \cdots & u_2 & l_n \end{pmatrix}. \tag{4.19}$$

通过初等行变换，矩阵 $(A,\boldsymbol{b})$ 等价于

$$\begin{pmatrix} 1 & 0 & 0 & \cdots & 0 & 0 & -\frac{u_2}{u_1} & -\frac{u_2}{u_1} & \cdots & -\frac{u_2}{u_1} & \frac{l_1-u_2}{u_1} \\ 0 & 1 & 0 & \cdots & 0 & 0 & \frac{u_2}{u_1} & 0 & \cdots & 0 & \frac{l_2}{u_1} \\ 0 & 0 & 1 & \cdots & 0 & 0 & 0 & \frac{u_2}{u_1} & \cdots & 0 & \frac{l_3}{u_1} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & 0 & 0 & 0 & \cdots & \frac{u_2}{u_1} & \frac{l_n}{u_1} \\ 0 & 0 & 0 & \cdots & 0 & 1 & 1 & 1 & \cdots & 1 & 1 \\ 0 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 & \cdots & 0 & 0 \end{pmatrix}, \tag{4.20}$$

我们可得到系数矩阵 A 的秩等于增广矩阵 $(A,\boldsymbol{b})$ 的秩，即 $n+1$.

注意到当 $n>1$ 时，$n+1<2n$. 因此，以上线性方程组有无穷多解。

□

最后，本文给出命题 1的证明：

证明. 首先，记两个概率向量 $\boldsymbol{s},\boldsymbol{p}$ 关于某个给定的 logit z_i 的 KD损失的导数为 ∂KD_i^p，同样可记它们的 LD损失为 ∂LD_i^p.

（1）根据引理1，线性组合 $\boldsymbol{l}=u_1\boldsymbol{p}+u_2\boldsymbol{q}$ 是良好定义的，且 $\boldsymbol{s},\boldsymbol{l}$ 关于某个 logit z_i 的 LD损失的导数为：

$$\begin{aligned} \partial LD_i^l &= s_{\tau i}-l_{\tau i} && (4.21)\\ &= u_1 s_{\tau i}+u_2 s_{\tau i}-(u_1 p_{\tau i}+u_2 q_{\tau i}) && (4.22)\\ &= u_1(s_{\tau i}-p_{\tau i})+u_2(s_{\tau i}-q_{\tau i}) && (4.23)\\ &= u_1\partial KD_i^p+u_2\partial KD_i^q && (4.24) \end{aligned}$$

（2）根据引理2，$\boldsymbol{l}$ 的分解存在，可被写为 $\boldsymbol{l}=u_1\boldsymbol{p}+u_2\boldsymbol{q}$.

那么公式 4.9仍然成立。 □

4.5.2 LD的梯度重缩放

作为 [292]的直接推论，对于两个近邻位置上的相对预测置信度，LD对于分布焦点损失（DFL） [115]保持了梯度重新缩放：

推论 1 设 $q_{\tau i}=p_{\tau i}+c_i+\eta_i$，其中 c_i 是教师模型在位置 i 的相对预测置信度，η_i 是一个均值为0的随机噪声，则 *LD*对 *DFL*的 *logit*梯度重缩放因子为

$$\mathbb{E}_\eta\left[\frac{\partial_i^{LD}}{\partial_i}\right]=\mathbb{E}_\eta\left[\frac{\sum_{s\neq i}\partial_s^{LD}}{\sum_{s\neq i}\partial_s}\right]=\gamma+\frac{\lambda}{\tau}\left(\frac{c_i}{u_i-p_i}\right), \tag{4.25}$$

其中 $\frac{\partial_i^{LD}}{\partial_i}$ 表示加上 *DFL*后的 *LD*损失关于某个 *logit* z_i 的梯度，u_i 与 u_j 是两个求和为1的常数，γ 与 λ 分别是 *DFL*与 *LD*损失的损失权重，τ 是 *LD*中的温度参数。

在证明该推论前，本文先给出 [292]中的引理，

引理 3 设 $q_{\tau t}=p_{\tau t}+c_t+\eta$，其中 c_t 是教师模型在位置 t 的相对预测置信度，η_i 是一个均值为0的随机噪声，则分类 *KD*对交叉熵损失的 *logit*梯度重缩放因子为：

$$\mathbb{E}_\eta\left[\frac{\partial_t^{KD}}{\partial_t}\right]=\mathbb{E}_\eta\left[\frac{\sum_{i\neq t}\partial_i^{KD}}{\sum_{i\neq t}\partial_i}\right]=\gamma+\frac{\lambda}{\tau}\left(\frac{c_t}{1-p_t}\right). \tag{4.26}$$

有了以上引理后，本文给出推论 1的证明：

证明.

根据文献 [115]，DFL被定义为两个分别在 i 位置与 j 位置上的交叉熵损失的线性组合，

$$\mathcal{L}_{DFL}=u_i\mathcal{H}(\boldsymbol{p},\boldsymbol{g^i})+u_j\mathcal{H}(\boldsymbol{p},\boldsymbol{g^j}), \tag{4.27}$$

其中 $\boldsymbol{g}^i=\{0,1\}^n$ 是真实标签，其在第 i 个位置上为1，其余位置为0.

DFL关于某个 logit z_i 的梯度为

$$\frac{\partial\mathcal{L}_{\text{DFL}}}{\partial z_i}=u_i(p_i-g_i)+u_jp_i=p_i-u_i, \tag{4.28}$$

并且我们仍沿用记号 ∂_i 来表示 $\frac{\partial\mathcal{L}_{\text{DFL}}}{\partial z_i}$.

使用 LD后，整个损失函数表达为：

$$\mathcal{L}^{LD}=\gamma(u_i\mathcal{H}(\boldsymbol{p},\boldsymbol{g^i})+u_j\mathcal{H}(\boldsymbol{p},\boldsymbol{g^j}))+\lambda\mathcal{H}(\boldsymbol{p}_\tau,\boldsymbol{q}_\tau), \tag{4.29}$$

加上 DFL后的 LD损失关于某个 logit $z_i \in \boldsymbol{z}_S$ 的梯度为：

$$\frac{\partial \mathcal{L}^{LD}}{\partial z_i} = \gamma u_i (p_i - g_i) + \gamma u_j p_i + \frac{\lambda}{\tau}(p_{\tau_i} - q_{\tau_i}), \tag{4.30}$$

并且我们仍记 $\partial_i^{LD} = \frac{\partial \mathcal{L}_{\mathrm{LD}}}{\partial z_i}$.

根据引理 3，我们有公式 4.30与公式 4.28的比值为：

$$\mathbb{E}_\eta \left[\frac{\partial_i^{LD}}{\partial_i} \right] = \gamma u_i \frac{p_i - g_i}{p_i - u_i} + \gamma \frac{u_j p_i}{p_i - u_i} - \frac{\lambda}{\tau} \frac{c_i}{p_i - u_i} \tag{4.31}$$

$$= \gamma + \frac{\lambda}{\tau} \frac{c_i}{u_i - p_i}. \tag{4.32}$$

而不正确的位置的梯度之和为：

$$\begin{aligned}
&\sum_{s \neq i} \partial_s^{LD} \\
&= \gamma u_i \sum_{s \neq i} p_s + \gamma u_j \sum_{s \neq i, j} p_s + \gamma u_j (p_j - g_j) + \tfrac{\lambda}{\tau} \sum_{s \neq i} (p_{\tau_s} - q_{\tau_s}) \\
&= \gamma u_i (g_i - p_i) + \gamma u_j (g_i - p_i) - \gamma u_j g_j + \tfrac{\lambda}{\tau}(q_{\tau_s} - p_{\tau_s}) \\
&= \gamma u_i (g_i - p_i) - \gamma u_j p_i + \tfrac{\lambda}{\tau}(q_{\tau_s} - p_{\tau_s}) \\
&= -\partial_i^{LD}.
\end{aligned} \tag{4.33}$$

类似可对 ∂_s 应用上述过程，证毕。 □

4.6 Logit模仿 *vs.* 特征模仿

由以上分析可以看到，LD与分类 KD享有等价的优化效果，二者联系颇为密切，它们分别对学生检测器的定位任务与分类任务的学习有着促进作用。这位本文提出的 LD以及分类 KD提供了一个统一的 logit模仿框架，如图所示。现在更进一步，这种在师生模型输出上施加的蒸馏（logit模仿）与在师生深层特征上施加的蒸馏（特征模仿）孰优孰劣仍是一个谜？并且这自然引出了一些有趣的问题：

- 关于检测性能方面，与特征模仿相比，logit模仿表现如何？特征模仿是否始终优于 logit模仿？
- 这两种不同的蒸馏技术有何特点？它们学习到的特征表示和 logit有何差异？

本节的内容将对上述问题进行探讨。

4.6.1 数值结果的定量比较

首先，本小节将提出的 LD与几种最先进的特征模仿方法进行比较。这里采用选择性区域蒸馏法，即对主蒸馏区域进行 KD和 LD，并对 VLR进行 LD。由于现代目标检测器通常配备有 FPN [105]，本文遵循先前的工作 [209–211]，重新实现他们的方法，并将所有特征模仿施加在 FPN输出特征上以进行公平比较。在这里，“FitNets” [206]表示对整个特征图进行蒸馏。“DeFeat” [210]意味着在真实框外的特征模仿损失权重大于真实框内部的权重。“Fine-Grained” [209]表示该方法在高质量的锚点位置上对深度特征进行蒸馏。“GI Imitation” [211]表示该方法根据学生和教师的鉴别性预测选择蒸馏区域。“Inside GT Box”表示在FPN输出特征上选择真实框内部的区域。“Main Region”表示在主蒸馏区域内进行特征模仿。

表 4.13: **Logit模仿 *vs.* 特征模仿.**“Ours”表示使用选择性区域蒸馏，即，“Main LD + VLR LD + Main KD”. 教师模型是 ResNet-101，学生模型是 ResNet-50. 结果在 MS COCO val2017上报告。

方法	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Baseline (GFocal [115])	40.1	58.2	43.1	23.3	44.4	52.5
FitNets [206]	40.7	58.6	44.0	23.7	44.4	53.2
Inside GT Box	40.7	58.6	44.2	23.1	44.5	53.5
Main Region	41.1	58.7	44.4	24.1	44.6	53.6
Fine-Grained [209]	41.1	58.8	44.8	23.3	45.4	53.1
DeFeat [210]	40.8	58.6	44.2	24.3	44.6	53.7
GI Imitation [211]	41.5	59.6	45.2	24.3	45.7	53.6
Ours	42.1	60.3	45.6	24.5	46.2	54.8
Ours + FitNets	42.1	59.9	45.7	25.0	46.3	54.4
Ours + Inside GT Box	42.2	60.0	45.9	24.3	46.3	55.0
Ours + Main Region	42.1	60.0	45.7	24.6	46.3	54.7
Ours + Fine-Grained	42.4	60.3	45.9	24.7	46.5	55.4
Ours + DeFeat	42.2	60.0	45.8	24.7	46.1	54.4
Ours + GI Imitation	42.4	60.3	46.2	25.0	46.6	54.5

从表 4.13中可以看出，FitNets对整个特征图进行蒸馏获得了+0.6 AP的增益。通过在 GT框外部设置更大蒸馏损失权重，DeFeat [210]的性能略优于在所有位置使用相同损失权重的情况。 Fine-Grained [209]关注 GT框附近的位置，产生了41.1 AP的结果，与使用 Main Region进行特征模仿的结果相当。 GI

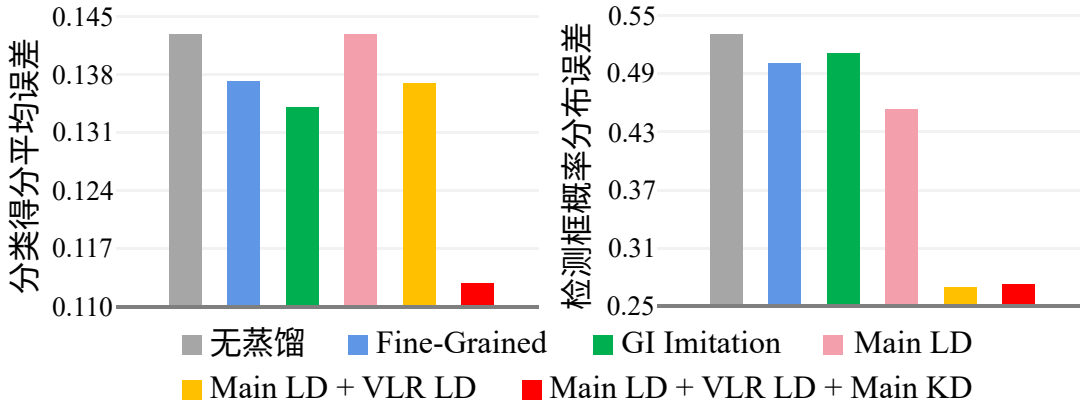


图 4.5: 先进的特征模仿方法与本文的 LD的视觉比较。柱状图展示了 P4、P5、P6和 P7 FPN层级上教师和学生之间分类分数和框概率分布的平均 L1误差。教师模型是 ResNet-101，学生模型是 ResNet-50。结果在 MS COCO val2017上进行评估。

Iimitation [211]寻找师生之间最具有区分性的区域而获得了41.5 AP。由于学生和教师之间预测存在很大差距，GI Iimitation的蒸馏区域可能出现在任何地方。尽管这些特征模仿方法有显著的改进，但它们并没有明确考虑知识分布模式。相反，本文的方法可以通过选择性区域蒸馏传递知识，直接获得42.1 AP的结果。值得注意的是，本文的方法是在 logit上操作而不是深度特征，这表明 LD是使logit模仿超越特征模仿的关键组成部分。此外，本文的方法与前面提到的特征模仿方法是可以相辅相成的。 表 4.13显示，使用这些特征模仿方法，本文方法的性能可以进一步提高。特别是，在使用 GI imitation的情况下，本文方法可将强大的 GFocal基线提高了+2.3 AP和+3.1 AP_{75}。

4.6.2 师生误差比较

接下来，本小节将检查分类得分和检测框概率分布的平均师生误差，如图4.5所示。可以看出，因为分类知识和定位知识混合在特征图上，Fine-Grained特征模仿 [209]和 GI imitation [211]按预期减少了这两个误差。“Main LD”和“Main LD + VLR LD”与 Fine-Grained特征模仿 [209]和 GI模仿 [211]相比，具有可比或更大的分类得分平均误差，但具有较低的框概率分布平均误差。这表明这两种设置仅使用 LD可以显著减少教师和学生之间的检测框概率分布距离，但不能减少分类头的误差。如果进一步在主蒸馏区域施加分类 KD，即得到“Main LD + VLR LD + Main KD”，则分类得分平均误差和检测框概率分布平均误差都可以减少。

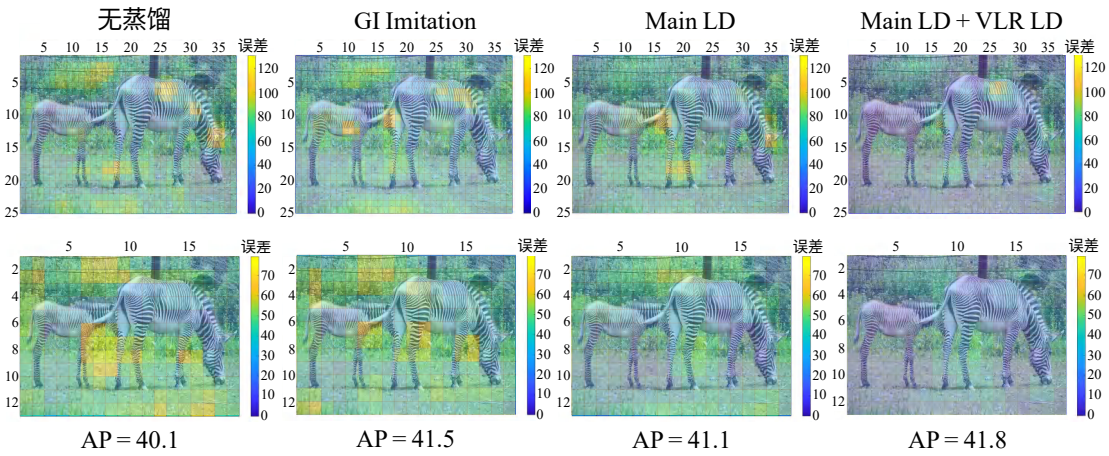


图 4.6: 先进的特征模仿方法与本文的 LD的视觉比较。该图展示了 P5（第一行）和 P6（第二行）FPN层级上教师和学生之间定位头 logit的 L1误差总和。教师模型是 ResNet-101，学生模型是 ResNet-50。最好在彩色图中查看。

本小节还可视化了 P5和 P6两个 FPN层级上学生和教师之间的定位头 logit的 L1误差总和。如图 4.6所示，与“无蒸馏”相比，GI imitation [211]确实减少了教师和学生之间的定位差异。需要注意的是，本文特意选择了一个精度表现略优于 GI imitation的模型（“Main LD + VLR LD”）进行可视化，但该方法仍明显减少了定位误差并缓解了定位的不确定性。

图 4.7分别绘制了学生和老师之间的平均误差，分别以深度特征、类别 logit和 bbox logit为基础。可以看出，这三种类型的错误在测试分辨率变化时表现出几乎一致的趋势。有趣的是，本文发现即使逻辑模仿可以缩小 bbox logit和分类逻辑的错误，它仍然学习到与老师完全不同的特征表示。从图 4.7的左图可以看出，本文方法增加了学生特征表示与教师特征表示之间的距离。此外，表 4.14显示 logit模仿在教师和学生的特征表示之间产生了几乎为零的皮尔逊相关系数。这表明，如果仅通过 logit模仿对学生模型进行训练，学生模型会产生与教师特征表示相距较远且非线性相关的特征表示。即便如此，logit模仿仍然可以获得表现良好的 logit输出以实现不错的泛化效果。根据表 4.14的最后一列和图 4.7的中图和右图表明，logit模仿不仅能使学生模型的输出 logit在距离上更接近教师模型的输出 logit，同时能在线性相关性上更接近。

4.6.3 AP景观

从特征级别或逻辑级别提取目标检测器是一个高维非凸优化问题，实际上容易但理论上困难。为了更好地理解逻辑模仿和特征模仿的行为，本小节提出

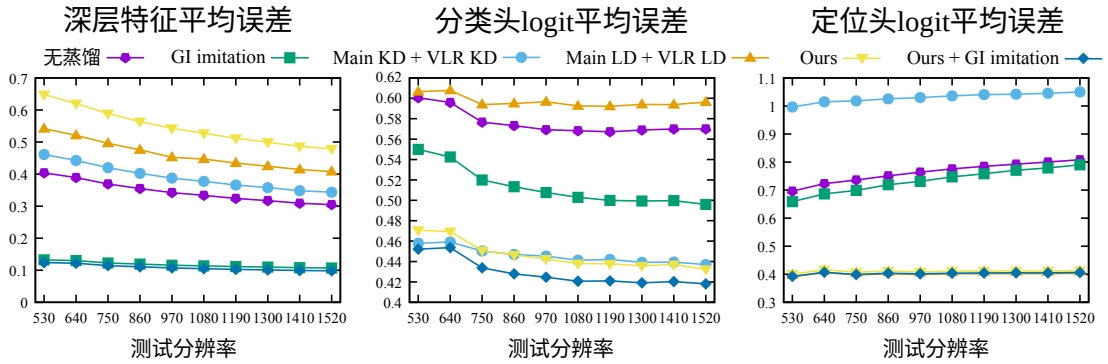


图 4.7: （左图）深度特征表示，（中图）分类头 logit和（右图）定位头 logit的平均师生误差。“Ours”表示“Main LD + VLR LD + Main KD”。曲线是在 MS COCO val2017上进行评估。

表 4.14: 师生之间的平均皮尔逊相关系数。“GI”：GI imitation [211]。“本文方法”表示“Main LD + VLR LD + Main KD”。结果在 MS COCO val2017上评估.

	无蒸馏	GI	本文方法	本文方法 + GI
深度特征	-0.0042	0.8175	-0.0031	0.8373
bbox logits	0.9222	0.9326	0.9733	0.9745

了一种新的可视化方法，称为 AP景观，专门用于目标检测，以观察学习特征表示中微小扰动引起的 AP变化。在 [293]中采用了一种经典的方法，该方法通过线性插值两个网络的参数来研究损失曲面的可视化。在 [293]中采用了一种经典的方法，该方法通过线性插值两个网络的参数来研究损失曲面的可视化。

在本文的可视化中，我们特别关注特征表示的经验性特征化以及它们如何影响最终的性能。考虑两个特征表示 M_f 和 M_l，它们是分别通过使用特征模仿和 logit模仿训练的学生检测器所学到的。本文分别展示了 AP景观的三维可视化（图 4.8第一行）及其在2D投影空间 $M_f \oplus M_l$ 上的可视化（图 4.8第二行）。本文使用两个标量参数 α 和 β，通过加权和 $M(\alpha,\beta) = \alpha M_f + \beta M_l$ 来获得一个新的特征表示。注意当 $\alpha = 0$ 且 $\beta = 1$ 时，这表示特征表示是由 logit模仿得到的，相反，当 $\alpha = 1$ 且 $\beta = 0$ 时，特征表示是由特征模仿得到的。然后，本文将 $M(\alpha,\beta)$ 输入到下游的分类头与定位头，并评估检测器的性能，绘制出最终的 AP分数。由于计算负担较重，本文将 $\alpha,\beta \in [-0.5, 1.5]$ 以可视化 AP景观。

从图 4.8中可以看出，logit模仿学习了稳健的特征表示，即红色五角星位于 $(0,1)$ 处，周围是一个平坦且表现良好的 AP分数区域。其次，本文观察到 GI imitation产生了比 logit模仿更为陡峭的 AP景观。本文将 GI imitation的 AP景观

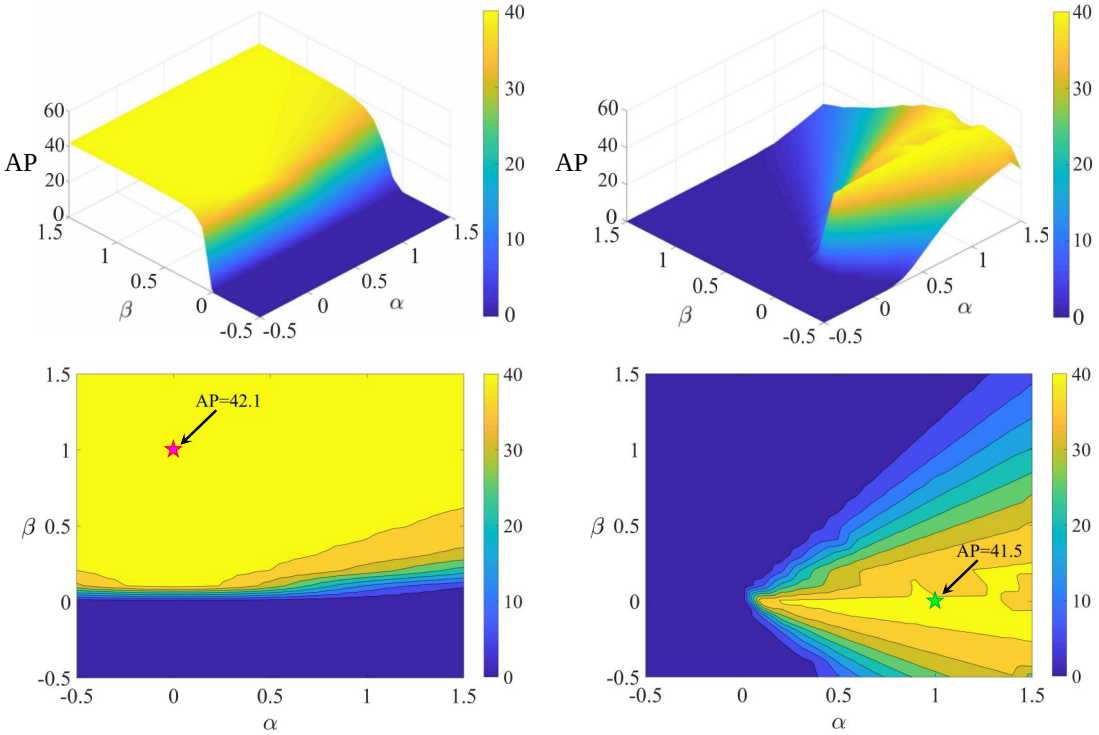


图 4.8: 特征子空间中的 AP景观。这些 AP景观由 MS COCO val2017上评估获得。

陡峭性归因于硬 l_2 损失监督。在这种情况下，对于学生来说，从教师那里模仿高级又先进的特征表示是非常困难的，因为教师是一个训练周期更长、精度更高的重型检测器。相反，logit模仿给予了特征表示更多的自由学习空间，从而实现更好的泛化能力。正如图 4.9所示，logit模仿还可以减少训练早期阶段的优化难度，而特征模仿在训练早期阶段的收敛速度较慢，泛化性能较差。

4.6.4 实验总结

基于以上的结果和观察，可以得出以下结论：

- 当明确进行定位知识蒸馏时，Logit模仿在目标检测中可以优于特征模仿。
- 特征模仿可以增加教师和学生之间特征表示的一致性，但会带来一些缺点，如特征的稳健性较差和训练收敛较慢。通过选择性区域蒸馏的Logit模仿可以显著提高教师和学生之间的 logit一致性，保持特征的学习自由度，从而加速训练过程并更有利于知识蒸馏性能提高。这表明，改进知识蒸馏性能的关键因素并不是教师和学生之间的特征一致性。

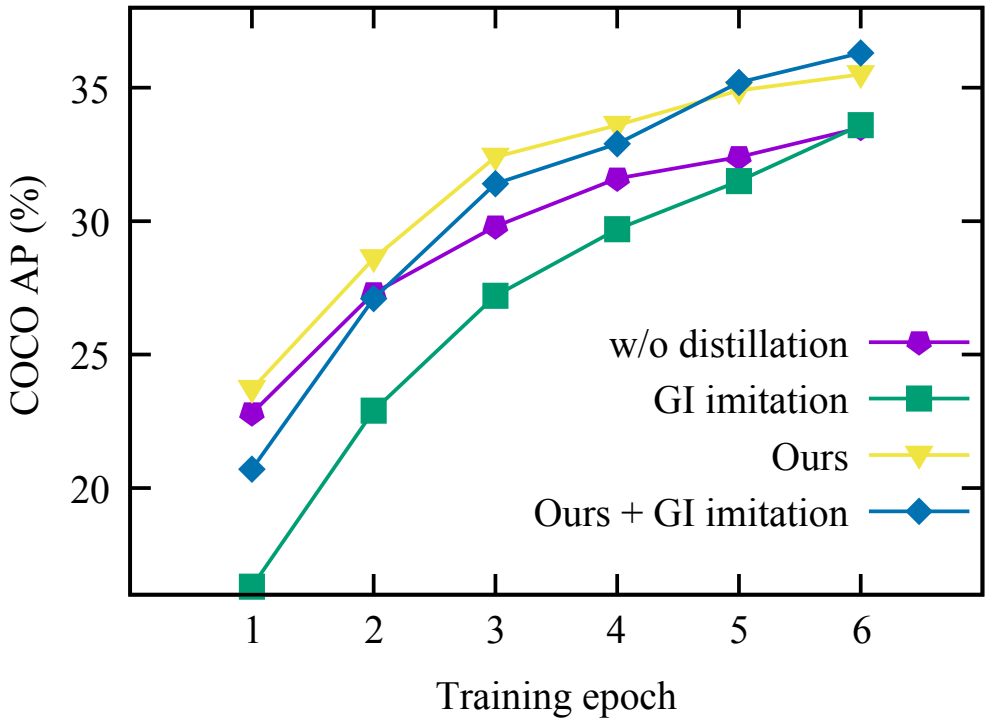


图 4.9: 训练早期阶段的平均精度（AP）。特征模仿显著减缓了收敛速度并导致了次优的泛化性能。Logit模仿（本文方法）可以减少训练早期阶段的训练难度。

4.7 本章小结

本章中首次提出了一种灵活的密集目标检测定位蒸馏方法，以及基于新的有价值定位区域的选择性区域蒸馏方法。本章展示了以下两点：1）对于目标检测，logit模仿可能比特征模仿更有效； 2）在进行目标检测蒸馏时，通过选择性区域蒸馏来传递分类和定位知识是重要的。本章的方法能够为目标检测领域提供新的研究启示，以便开发更好的蒸馏策略。在未来，将 LD方法应用于稀疏目标检测器（如 DETR系列 [117]），异构目标检测器组合，以及其他相关领域，例如实例分割、目标跟踪和三维目标检测，都值得进一步研究。此外，由于本文的 LD方法在优化效果上与分类蒸馏方法相当，一些改进的蒸馏方法可能也能够为 LD带来增益，例如关系蒸馏 [215]、自蒸馏 [281, 282]、教师助理蒸馏 [294]和解耦蒸馏 [295]等。跨架构蒸馏利用最近的先进分类模型作为教师模型（如 [269, 296–299]）也是一个有趣的探索方向。

第五章 SlimHead：高效能紧致表达的检测头网络

得益于多层级学习框架的成功，密集目标检测多年来一直很受欢迎。在常见的检测头中，通过将各种尺度的物体的学习分发至多层级特征金字塔，这种分而治之的解决方案减轻了模型优化难度。然而，一个普遍被忽视的问题是，目标检测的检测头网络的紧致表达仍然不足。由于浅层特征图的分辨率较高，这导致了检测头的计算量变得十分庞大，严重减慢了模型的推理速度。目标检测的效能严重受阻。为了解决这个问题，本文通过研究性能敏感性来探索多层级头部网络设计。其探索成果是一种新型检测头网络，称为 SlimHead，一个简单、高效且易于推广的检测头网络，进一步释放了密集目标检测器的多层级学习的潜力。该方法的设计理念是在检测头网络之前和之后为浅层特征注入两个插值函数来实现加速，同时保持了相近的检测精度。得益于其灵活性，该方法可以轻松集成具有更高计算复杂度的基础算子操作以提高检测精度，而不会损失推理效率。 SlimHead适用于多种高层视觉任务，例如有向目标检测、行人检测和实例分割。最后，在 PASCAL VOC、MS COCO、DOTA和 CrowdHuman上进行的大量实验证明了本章方法的广泛适用性和很高的实用价值。

5.1 引言

密集目标检测是计算机视觉领域的一个长期研究课题，并持续对相关领域产生积极影响，例如有向目标检测 [77, 300]、行人检测 [245]和实例分割 [301, 302]等等。截至目前，凭借其优异的速度与精度的权衡以及对低端边缘设备的友好性，密集目标检测在工业应用中仍然占据着不可撼动的主导地位。物体有大有小。因此，学界提出了一种多层级学习的解决范式，将大物体的学习交给深层特征图（低分辨率），而将小物体的学习交给浅层特征图（高分辨率）。 FPN [105]，中文名称为“特征金字塔网络”，便是该范式最具代表性的方法之一。其构建了一个多层级特征金字塔来处理主干网络特征并在多个层级上并行地进行实例预测任务。这种学习范式已被证明是有效的，并因此在密集目标检测领域占据主导地位 [101, 104, 106, 303–305]。

最近，密集目标检测领域的一系列研究突破主要集中在增强分类和定位

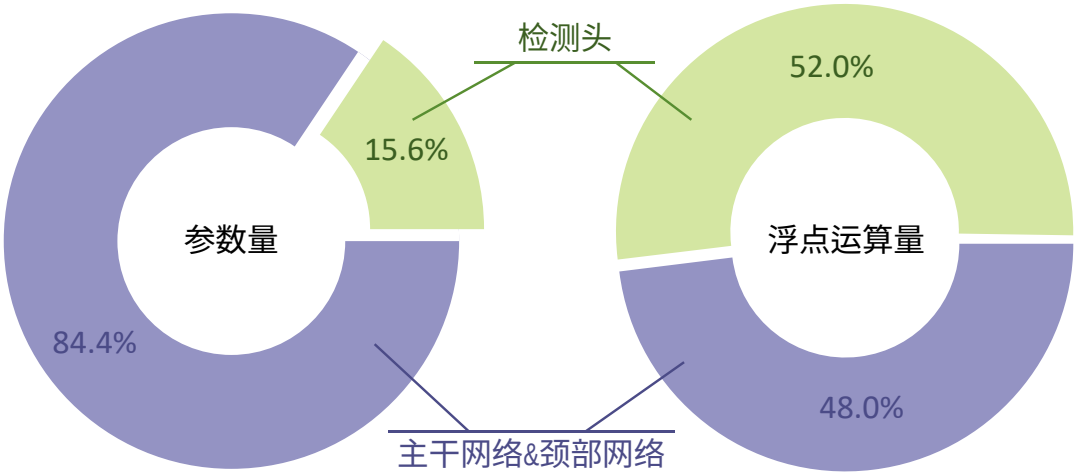


图 5.1: 尽管检测头网络在模型参数量方面很轻量，但它们占据了很大一部分计算复杂度。这个问题在密集目标检测器中很常见。

之间的一致性 [115, 130–132, 202]、缓解定位模糊性 [115, 196, 197, 241, 242]、以及提高定位质量 [200, 201, 239, 306]等等。尽管多层级学习是上述所有工作的基础，但仍然存在一个常常被忽视的问题，并且很少受到关注： *高分辨率下的浅层级太耗时了*。 该缺陷导致的问题是，尽管检测头网络在模型参数上是轻量级的，但它们却占用了相当大的计算量。如图 5.1所示，仅具有15.6%参数量的检测头网络就可以产生多达52.0%的浮点运算量（FLOPs）。此外，检测头网络对特征的处理通常在各层级上是并行的，这意味着检测头中的基础算子操作在不同层级之间共享。这使得那些有利于提高检测精度的重型算子（例如可变形卷积 [284]）一旦集成进检测头中，那么模型计算量将更加繁重。

本章节将通过探究目标检测器的性能敏感性来重新思考多层级学习范式。深入研究检测头网络的本质属性并找出哪些组件对目标检测器的性能和效率至关重要，从而试图找到一种高效能的解决方案。通过对检测头网络进行性能敏感性分析，本章节总结了检测头网络的本质属性在于细化特征与定义解空间。作为探究的成果，本章节得到了一个非常简单、高效且易于推广的检测头网络，称为 SlimHead。其设计理念分两个阶段：瘦身和增重。在第一个瘦身阶段中，本文在检测头网络发挥作用之前注入一个插值函数来对特征金字塔进行“瘦身”。这将产生一个紧致而高效的检测头网络。在第二阶段“增重”中，本文使用逆插值函数来对特征金字塔进行“增重”，从而使特征恢复到了原始解空间。本文发现，上述操作将成为 SlimHead与传统检测头网络之间的关键区别，并对

于减少计算量和保持检测精度至关重要。实验表明，当正确集成这两个阶段的操作时，这种即插即用的策略可以优雅地对齐预测的解空间，而不需要再进行额外的修改。结果是，SlimHead使我们能够显著减轻检测头网络的计算负担，同时保持相当的检测精度。更进一步，计算复杂度更高的基础算子操作（例如可变形卷积 [284]）可以被轻松集成进来，实现在不损失效率的情况下提高准确度。作为一项额外的好处，SlimHead还可以节省 GPU内存的占用量，例如在ResNet-18的模型上减少了15.1%，这对于在低端边缘设备上的部署更加友好。

本章节的亮点有如下两点：

1. 本章节重新思考了密集目标检测中的效率瓶颈问题。本章节探索的结果是一个非常简单、高效且易于推广的检测头网络，称为 SlimHead。它实现了更好的速度-精度权衡。由于其灵活性，更高计算复杂度的基础算子可以毫不费力地集成进来，从而实现检测精度提升而不会降低运算效率。SlimHead的优势：*性能更高、速度更快、易于实现、GPU内存占用更低*。
2. 本章节还将 SlimHead扩展到多个高层视觉任务，例如有向目标检测、行人检测和实例分割。在 PASCAL VOC [36]、MS COCO [37]、DOTA [29]和CrowdHuman [245]上的大量实验证明了该方法具有广泛的适用性和很高的实用价值。

5.2 多层级学习范式及分析

多层级学习可被定义为一个并行优化问题：

$$\min_{\Theta} \sum_{i} \mathcal{L}(\mathcal{H}(X_i|\Theta), G_i), \tag{5.1}$$

其中 i 是层级序号，X_i 是第 i 层级输入特征图, $\mathcal{H}$ 是检测头网络并具有网络参数 Θ. G_i 是第 i 层级上的真实值，用于提供监督信号。 $\mathcal{L}$ 是给定的损失函数。在密集目标检测领域，多层级学习的一个形式可参见图 5.2，其中共含有5个特征层级，从浅到深依次为 P3，P4，P5，P6，P7，这在许多经典的目标检测器中被采用，例如 RetinaNet [106], FCOS [101], GFocal [115], TOOD [131]等等。不难看出，对于这样的优化问题，有以下三个因素可能影响着检测模型的性能与效率。

1. 检测头网络 $\mathcal{H}$ - 将语义特征映射到具有特定物理意义的 logit向量。一般来说，$\mathcal{H}$ 由一系列堆叠卷积组成，以起到细化和再加工 FPN输出特征的作用。值得注意的是，由于检测头网络参数是共享的，$\mathcal{H}$ 通常在不同级别上

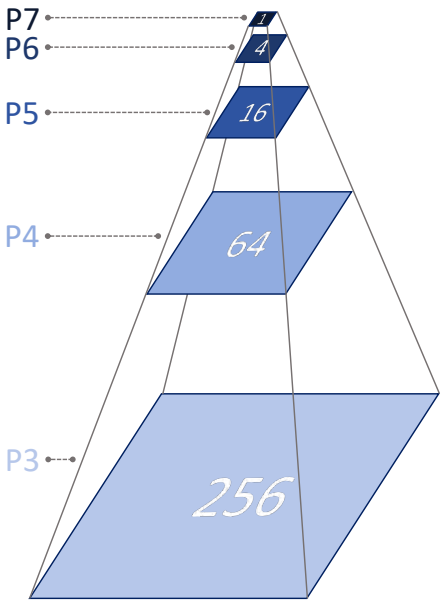


图 5.2: 特征金字塔的比例尺寸。检测头网络在浅层级上占用了大量的计算量。每一层级的计算量都是较深一层级的四倍。

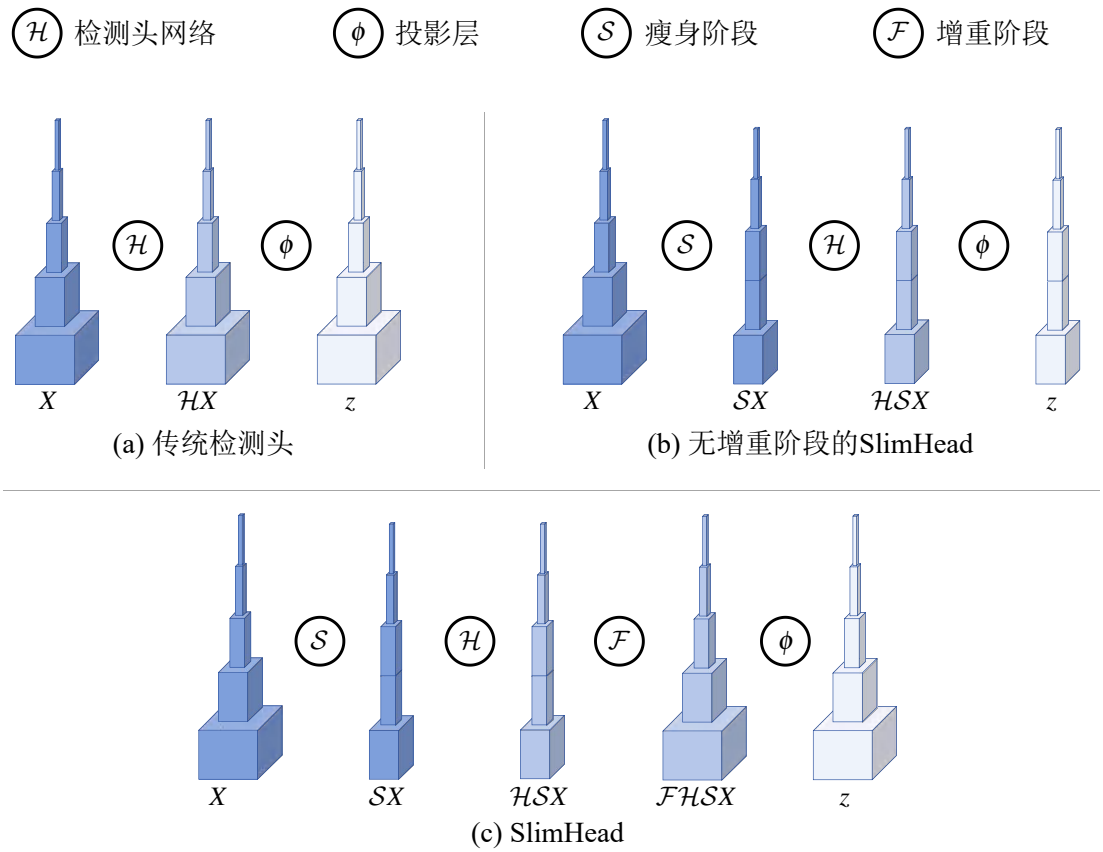
并行处理每一层级特征。

2. 输入特征 X_i - 影响整个优化过程和模型效率。主干网络+颈部网络越强，所获得的特征就越稳健。
3. 损失函数 $\mathcal{L}$ - 决定优化方向。这在各个层级之间也是并行的。

接下来，让我们更深入地观察输入特征 X_i. 首先，$X_i \in \mathbb{R}^{b \times C \times H_i \times W_i}$ 是上游网络的输出特征，即主干网络+颈部网络，其中 b 与 C 是批量大小与通道数，W_i 与 H_i 是特征图的宽与高。X_i 的分辨率将决定检测头的推理速度。X_i 的分辨率越大，检测头的推理速度就会越慢，反之，分辨率越小，推理速度就越快。第二，X_i 的分辨率同时还将决定锚框的设定个数与尺寸，这对于模型精度有着重要影响。由于损失函数 $\mathcal{L}$ 与模型的推理速度无关，下面将对检测头网络 $\mathcal{H}$ 与输入特征 X_i 进行性能敏感性分析，以观察它们对于检测模型性能与效率的影响。

5.3 SlimHead方法介绍与分析

本节的目标是寻找构建高效、强大的检测头网络所必需的组件。通过对检测头网络的性能进行敏感性分析，最终，一个简单、高效、灵活、易于推广的检测头网络自然诞生。



(a) 传统检测头　(b) 无增重阶段的SlimHead

(c) SlimHead

图 5.3: 多层级学习的三种检测头网络设计。 X：上游网络输出特征，如主干网络+FPN。 z：分类分支与定位分支的输出 logit。 (a) 传统检测头由5个特征层级构成，由浅至深为 P3至 P7。 (b) 无增重阶段的 SlimHead：仅在检测头网络处理特征前使用瘦身阶段，这将释放浅层的计算负担。 (c) SlimHead：进一步配置了增重阶段，使输出特征 $\mathcal{H}\mathcal{S}X$ 恢复至原始分辨率

5.3.1 传统检测头分析

传统检测头是现有的密集目标检测器的基础组成模块 [101, 106, 114, 115, 129–132]，如图 5.3 (1) 所示。给定检测头网络 $\mathcal{H}$ 与一个投影层 ϕ，由上游网络所输出的多层级特征 $X = \{X_i\}, i = 3,4,5,6,7$ 将被 $\mathcal{H}$ 所细化，并投影至输出 logit $z = \phi\mathcal{H}X$. 减少检测头网络计算负担的一个简单方法是减少卷积层的数量。为探究该方法的效果，本文逐一减少检测头网络中所含的卷积层数量。其结果由图 5.4展示，可以看到检测精度随着卷积层数量的减少而降低，尽管模型推理速度在提升。这表明检测头网络需要多个卷积算子操作来细化特征，而使用较少的卷积算子操作无法达到良好的检测精度。那么，检测头网络中的哪些组件对于实现更好的速度-精度权衡是至关重要的呢？

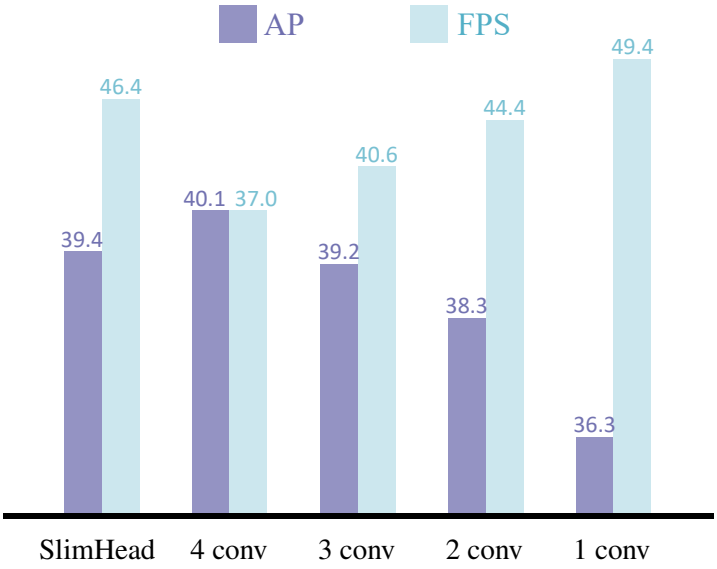


图 5.4: 检测精度（AP）与速度（FPS）关于检测头网络中卷积层数的变化趋势

5.3.2 SlimHead设计

根据图 5.2，浅层级的计算过于耗时，极大地限制了密集目标检测器的效率。在上一小节中，通过对传统检测头的分析，本文展示了检测头网络的重要性，并且足够的卷积算子操作是取得良好性能的关键。本文发出疑问：*多层级学习能否在保持相当的检测性能的情况下拥有更快的推理速度？* 在本节中，我们将鼓励探索检测头网络的本质属性，并充分利用这些新发现来构建出一个速度-精度权衡更好的目标检测器。为了使行文流畅，下面首先介绍本文所提出的SlimHead，其设计理念分为两个阶段：“瘦身”和“增重”。

阶段-I：瘦身：本文提出在检测头网络 $\mathcal{H}$ 处理特征前，插入一个带有缩放因子 r 的插值函数 $\mathcal{S}$. 注意到，这在概念上可以应用于任何层级，但本文发现对于深层级来说这是不必要的，因为它们不是模型的效率瓶颈。因此，本文引入一个层级选择参数 $K \in \{3,4,5,6,7\}$ 来选择 $i \leqslant K$ 的层级来应用瘦身阶段：

$$\mathcal{S}X = \mathcal{S}(X_i; r), \quad i \leqslant K. \tag{5.2}$$

阶段-II：增重：在这一阶段，本文进一步在检测头网络之后、投影层之前，插入一个带有缩放因子 $1/r$ 的插值函数 $\mathcal{F}$，表达如下：

$$\mathcal{F}\mathcal{H}\mathcal{S}X = \mathcal{F}\left(\mathcal{H}(\mathcal{S}(X_i; r)); \frac{1}{r}\right), \quad i \leqslant K. \tag{5.3}$$

这一操作使得被瘦身的特征重新转换回了它们原始的分辨率，这保证了每个锚点上预测框的数量不变。所有与训练过程有关的超参数将保持不变，例如相同的锚框定义，相同的标签分配，以及相同的损失函数超参数。因而优化的解空间保持不变。 SlimHead的示意图可见图 5.3 (c)。从形象上来看，瘦身阶段后，检测头上的特征金字塔变得更为纤细，而在增重阶段后，特征金字塔重回了臃肿形态。

5.3.3 检测头网络的本质属性

现在，本小节来回答之前提到过的疑问。本文首先比较了两种 SlimHead设计的性能与效率。第一种为无增重阶段的 SlimHead，如图 5.3 (b) 所示。第二种为 SlimHead完全体，如图 5.3 (c) 所示。在图 5.5中，本文展示了上述两种SlimHead设计的性能敏感性。可以看到这两种 SlimHead设计都能够大幅降低目标检测器的运算复杂度。检测头网络的浮点运算量被显著降低。然而有趣的是，无增重阶段的 SlimHead却显示了严重的检测精度下降，约下降了7.5的 AP值。而如果重新添加增重阶段，SlimHead完全体则可以取得39.4的 AP，这与原始基线模型的40.1 AP相当，并且收获了25.4%的推理速度提升。这表明由于标签分配和损失函数中涉及的所有超参数都是根据输出 logit图的大小而定制的，因此保持优化的解空间不变是非常有必要的。 图 5.4显示，SlimHead取得了非常有希望的检测结果，实现了更好的速度-精度权衡。

根据以上实验结果，本文可以获得总结，检测头网络实际上由下列两个本质属性构成：

1. 细化特征：进一步细化上游网络（FPN）的输出特征。
2. 定义解空间：决定了预测输出的维度与数量，并定义了包含标签分配和损失函数在内的所有超参数。

这为我们进一步提高多层级学习的效率提供了机会：只要 *logit*图分辨率保持不变，理论上就可以缩小特征图分辨率以加快推理速度，同时保持相当的检测精度。

5.3.4 SlimHead的特性

所提出的 SlimHead有以下四个吸引人的特性：

1. 瘦身阶段使得特征金字塔 X 变得紧致，大大减轻了浅层的计算负担。当缩放因子 $r = 1$ 时，SlimHead退化为原始检测头网络。而当 $r < 1$ 时，被瘦身

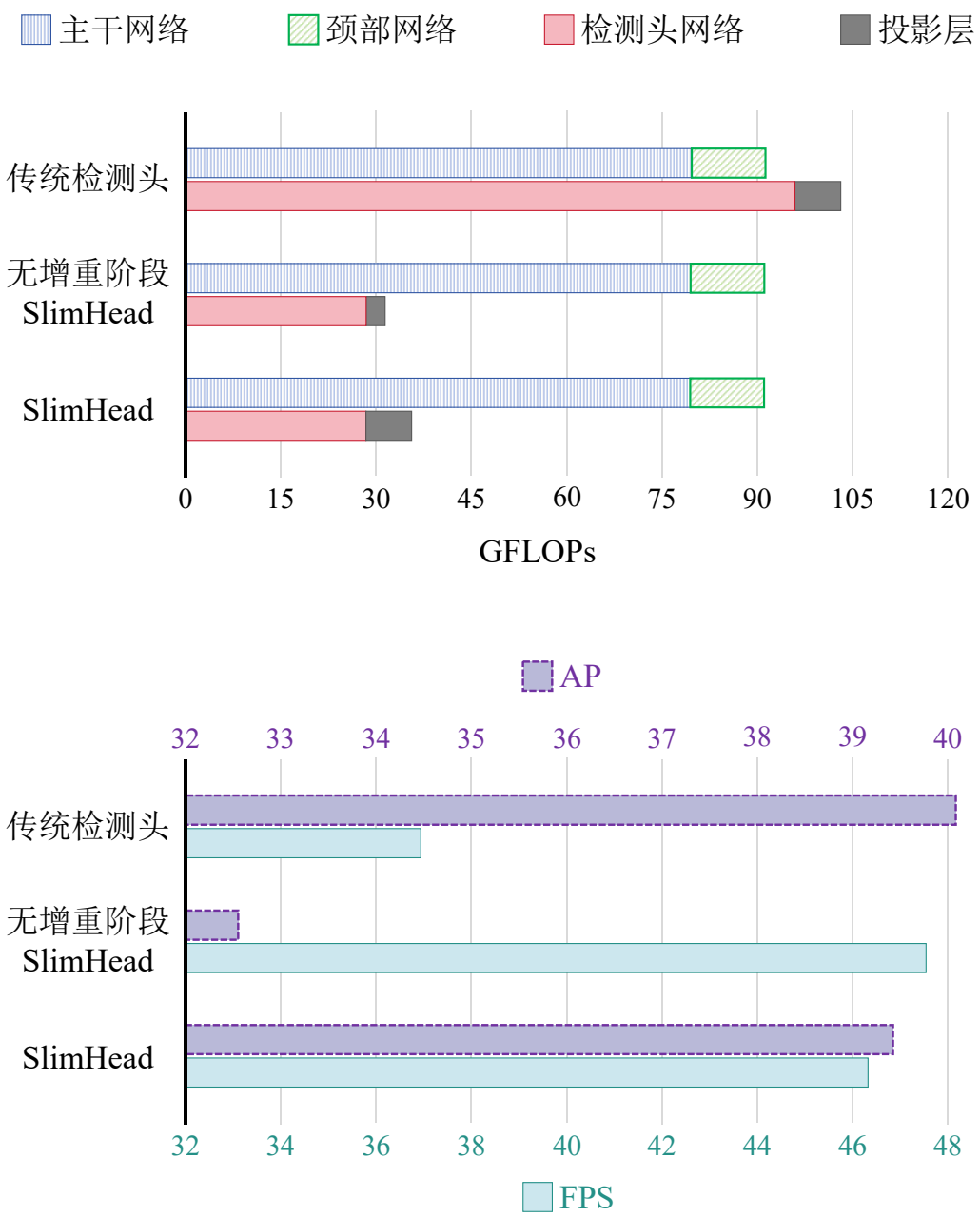


图 5.5: 三种检测头设计的浮点运算律（GFLOPs）、检测精度（AP）、推理速度（FPS）的比较。SlimHead可以显著降低检测头网络的运算量，提高25.4%的推理速度，同时检测精度相当。同时，无增重阶段的 SlimHead会导致严重的检测精度下降。

的层级的计算复杂度降低为原来的 r^2。

2. 由于计算复杂度降低，SlimHead允许集成计算复杂度更高的卷积算子操作，而不会严重降低速度，例如可变形卷积 DCN [284]。

3. 随着浅层特征图分辨率的降低，GPU内存占用量也将显著减少。
4. SlimHead具有很强的通用性。在大多数以前的目标检测方法中 [101, 106, 115, 131]，无论它们使用何种特征聚合操作，SlimHead都可以纳入其中。

5.4 消融实验

本节采用 MMDetection [265]框架进行消融实验，并使用 MS COCO [37]数据集进行验证。除了缩放因子 r 和层级选择参数 K 之外，所有超参数均保持不变，以便进行公平的比较。在消融研究中，本节采用流行的单阶段目标检测器 GFocal [115]为基线模型，其配置了 ResNet-50 [108]主干网络和 FPN [105]颈部网络。除非另有说明，否则本节默认采用经典的单尺度 1×（12个时期）训练计划，训练分辨率为 1333×800。 2×（24个时期）训练计划表示以多尺度图像分辨率进行训练，其中图像短边的随机采样范围为 $[480 : 960]$。本节以 MS COCO的平均精度（AP）为主要报告指标。由于本章方法是为了实现高效的目标检测而提出的，本节还报告了每秒运行帧数（FPS）以评估模型的推理速度。所有目标检测器的推理速度都是在单块 RTX 3090显卡上测量的。

插值函数 $\mathcal{S}$

这里，本文研究了3种不同类型的插值函数 $\mathcal{S}$。第一个是最近邻插值。第二个是最大池化算法。第三个是双线性插值。在这个实验中，本文只在最浅的层级上应用插值函数 $\mathcal{S}$，即层级选择参数 $K = 3$. 结果报告在表 5.1中。可以看出，无论使用哪种插值算法，本文的方法都可以达到与基线模型相当的检测精度，同时计算效率更高。在这3个插值算法中，最近邻插值的一个最简单、最高效，也达到了最高的检测精度。因此，在接下来的实验中，本文默认采用最近邻插值算法。

表 5.1: **不同的插值函数 $\mathcal{S}$**：结果报告在 MS COCO val2017数据集上。

插值算法	AP	AP_{50}	AP_{75}	FPS
基线（无插值）	40.1	58.2	43.1	37.0
最近邻插值	**39.7**	57.8	**42.8**	**43.9**
最大池化算法	39.7	**58.0**	42.7	43.2
双线性插值	39.5	57.8	42.5	43.7

层级选择参数 K

SlimHead的核心在概念上可以应用于任何层级。这里本文对层级选择参数 K 进行了一项实验，观察将 SlimHead应用于从浅到深的特征层级时检测性能的变化。得益于 SlimHead的特性2（参见章节 5.3.4），本文在检测头网络的前两层卷积上应用可形变卷积 DCN [284]。 表 5.2表明，当 $K \leqslant 5$ 时，即SlimHead仅应用于浅层级时，本文的方法可以实现更好、更快的性能。本文发现将 SlimHead应用于更深的层级后，检测性能反而会下降。此外，SlimHead应用于较深的层级对推理速度没有太多好处，因为深层级的特征图尺寸本身较小，并非模型计算复杂度的瓶颈。因此，在实践中，本文通常仅在浅层级应用SlimHead。

表 5.2: **不同的层级选择参数 K**：结果报告在 MS COCO val2017数据集上。

K	AP	AP_{50}	AP_{75}	FPS
基线	40.1	58.2	43.1	37.0
3	**41.4**	**59.4**	**44.8**	38.0
4	41.1	59.2	44.3	40.7
5	40.6	58.9	43.7	41.3
6	40.0	58.5	43.4	41.3
7	39.6	58.2	42.9	**41.9**

缩放因子 r

在应用 SlimHead时，浅层级特征将被临时转换至较小的特征空间，以降低计算复杂度。这里，本文研究了缩放因子 r 的影响，结果报告在表 5.3中。在这个实验中，层级选择参数 $K = 3$，即只将 SlimHead应用于最浅的层级。从表 5.3中的第一组可以看出，本文的方法可以达到与基线模型（$r = 1$）相当的检测精度。特别是，当 $r = 0.5$ 时，SlimHead实现了43.9的 FPS，将检测器的推理速度加快了近 20.3%。更进一步，若将 DCN [284]合并到本文的 SlimHead中，其中只替换了检测头网络的前两层卷积算子。 表 5.3中的第二组实验结果展示了这种改变可以带来检测精度的显著提升。而如果本文直接在原始检测头网络（$r = 1$）上使用 DCN替换普通卷积，目标检测器的速度将显著降低至29.9FPS，这是因为浅层级占用了相当大的计算负担，因此对于此类改进的卷积算子操作，直接替换所带来的额外计算负担是无法承受的。而本文的 SlimHead在 $r = 0.5$ 时

实现了41.4 AP和38.0 FPS，这甚至比基线模型（第1行）更快更好。这表明本文的方法可以将目标检测器的 AP提高1.3，同时获得免费的推理速度加速。此外，值得注意的是，本文的方法在 $r = 0.5$ 时实现了 AP提升的峰值。这表明最好应该将浅层级特征转换为与之邻近的层级大小，这可能会带来相似的梯度流，二者处于相同大小的维度空间中。因而在其它所有实验中，本文默认设置缩放因子 $r = 0.5$.

表 5.3: **不同的缩放因子** r：结果报告在 MS COCO val2017数据集上。

r	DCN	AP	AP_{50}	AP_{75}	FPS
1.0		**40.1**	**58.2**	**43.1**	37.0
0.9		39.2	57.7	42.4	38.0
0.8		39.2	57.6	42.1	39.9
0.7		39.6	**58.0**	42.6	40.6
0.6		39.2	57.6	42.4	42.2
0.5		**39.7**	57.8	**42.8**	43.9
0.4		37.9	56.3	40.2	**45.1**
1.0	✓	**42.0**	**60.0**	**45.6**	29.9
0.9	✓	40.9	59.3	44.3	31.7
0.8	✓	40.9	58.9	44.2	34.0
0.7	✓	41.0	59.2	44.4	35.2
0.6	✓	40.7	58.9	44.1	36.8
0.5	✓	**41.4**	**59.4**	**44.8**	38.0
0.4	✓	39.6	57.9	42.3	**39.6**

SlimHead对不同尺度物体的影响

由于 SlimHead改变了检测头网络的特征图大小，因此它可能对不同尺度的物体产生不同的影响。首先，表 5.4报告了小、中、大物体的平均精度 AP_S、AP_M 和 AP_L。可以看出，SlimHead提高了中物体和大物体的 AP性能，但在小物体上显示出轻微的 AP下降。这可能是因为 SlimHead保持了中层级和深层级的特征不变，而浅层级配备了 SlimHead。这使得浅层级的信息丢失不可避免地导致了小物体一定程度的性能下降。尽管如此，本文的方法大大提高了中物体/大物体的检测性能。

此外，本文接着对各种物体尺度的准确率进行了更全面的评估。这里本文沿用了第四章所提出的区域评估的思想，将评估的物体尺度设置为

表 5.4: **SlimHead对不同尺度物体的检测性能**：结果报告在 MS COCO val2017数据集上。

SlimHead	FPS	AP	AP_S	AP_M	AP_L
	37.0	40.1	23.3	44.4	52.5
✓	38.0	41.4	22.7	45.4	55.9

$t = 0, 16, 32 \cdots, 320$。如果真实框和检测框的面积落在区间 $[t^2, (t+16)^2]$ 内，则会将它们筛选出来评估检测器。如图 5.6(a)所示，SlimHead可以在 32^2 的物体尺度上实现相当的 AP性能。而在非常小的物体尺度 $t \leqslant 16$（即微小物体）下降低了 AP。在图 5.6(b)中，可以看到当物体尺度 $t > 32$ 时，本文的方法显示出明显的优势。当 $t < 32$ 时，由于性能下降幅度很小，因此两条 AP曲线紧密贴合。这表明 SlimHead在很大的物体尺度范围内产生了积极的影响。

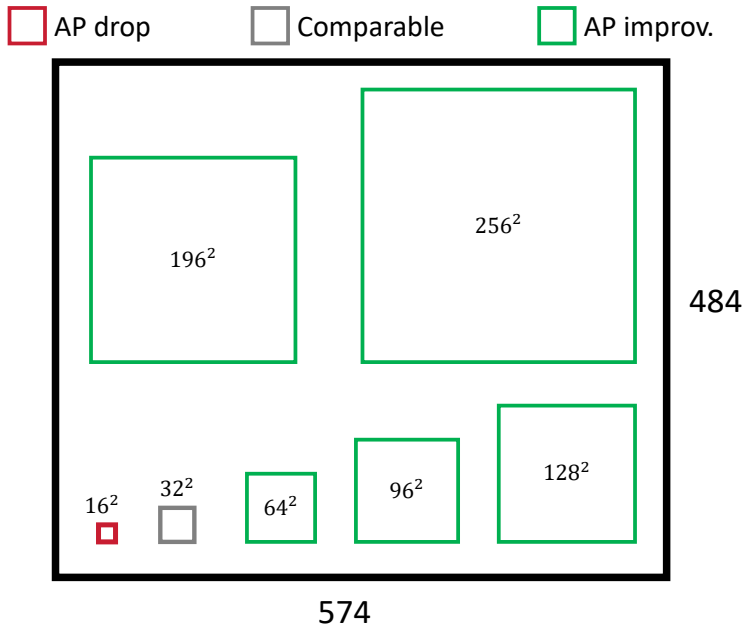
5.5 SlimHead的推广应用

本节将对 SlimHead推广应用，以验证其通用性，探究其在不同场景下的适应能力。本节将对4个流行的目标检测数据集进行实验，分别是 PASCAL VOC [36]、MS COCO [37]、DOTA-v1.0 [29]、以及 CrowdHuman [245]，关于它们的详情介绍可参见章节 2.5。此外，本节还将对3个高层视觉任务进行实验，包括通用目标检测、有向目标检测、实例分割。

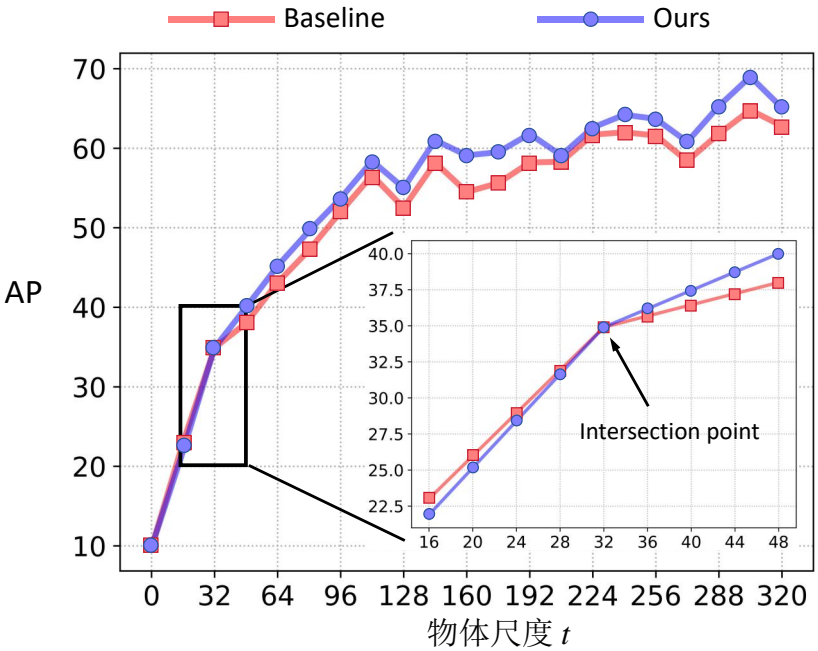
SlimHead纳入密集目标检测

首先，SlimHead可以很方便整合到现有的密集目标检测器中。这里，本文选择了一些代表性的密集目标检测器，它们均是构成当前最先进的目标检测器的基石。第一个是经典的多阶段密集到稀疏目标检测器 Faster R-CNN [96]。本文在 RPN上应用 SlimHead，因为 RPN可以被视为具有二分类的密集目标检测器。第二个是 FCOS [101]以及其后续改进版本 ATSS [114]、GFocal [115]、TOOD [131]。 DCN将应用于所有这些目标检测器以提高检测精度，并且本文保持应用 SlimHead后的推理速度比原始检测器略快一些。

训练采用单尺度 1×（12个 epoch）训练计划，这是检测社区中的经典训练设置。结果在表 5.5中报告。可以看出，本文的 SlimHead提高了所有5个目标检测器的检测精度。在 Faster R-CNN、FCOS、ATSS、GFocal和 TOOD上，AP分别提高了 +0.4、+0.5、+1.2、+1.3、+0.9，AP_{75} 提高了+0.9、+0.3、+1.2、+1.7、



(a) SlimHead对不同尺度物体的影响



(b) 各种尺度范围物体的 AP曲线

图 5.6: (a) SlimHead对不同尺度物体的影响。图像的平均尺寸以黑框展示。 (b) 各种尺度范围物体的 AP曲线。评估针对框面积介于区间 $[t^2,(t+16)^2]$ 的物体。

+1.1。重要的是，SlimHead甚至带来了推理加速。这表明本文的方法可以在密集目标检测中实现更好的速度-精度权衡。

表 5.5: **SlimHead在5个密集目标检测器上的检测性能**：结果在 MS COCO val2017上报告。FPS是在单块 RTX 3090显卡上测量的。

目标检测器	SlimHead	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	FPS
Faster R-CNN [96]		37.4	58.1	40.4	21.2	41.0	48.1	37.7
	✓	37.8	58.7	41.3	21.6	41.5	49.1	38.7
FCOS [101]		38.7	57.4	41.8	22.9	42.5	50.1	36.9
	✓	39.2	57.9	42.1	21.8	43.1	52.5	38.0
ATSS [114]		39.3	57.5	42.8	24.3	43.3	51.3	36.9
	✓	40.5	58.4	44.0	22.9	44.5	53.8	38.0
GFocal [115]		40.1	58.2	43.1	23.3	44.4	52.5	37.0
	✓	41.4	59.4	44.8	22.7	45.4	55.9	38.0
TOOD [131]		42.3	59.6	45.9	25.8	45.6	54.9	33.6
	✓	43.2	60.4	47.0	24.2	47.0	58.5	34.2

SlimHead在不同检测数据集上的效果

这里，本文进一步通过在更多检测数据集上进行实验来检查 SlimHead的通用性，包括 PASCAL VOC [36]、CrowdHuman [245]和 DOTA [29]。这些数据集的训练/评估集划分可以在章节 2.5中查阅。对于 PASCAL VOC，本文训练检测器12个 epochs，在9个 epoch后学习率降低为0.1倍。对于行人检测数据集 CrowdHuman，本文训练检测器30个 epochs，在第24个和第28个 epoch后学习率乘以0.1。对于 VOC和 CrowdHuman，本文采用水平框密集目标检测器 TOOD [131]，以 ResNet-50为主干网络和以 FPN为颈部网络。对于 DOTA，本文采用具有相同 ResNet-50主干网络和 FPN颈部网络的遥感领域经典有向密集目标检测器 PSC [300]。

如表 5.6所示，SlimHead在保持高效率的同时，一致地提升了三个检测数据集上的检测性能，证明了该方法的良好泛化能力。具体来说，SlimHead在 PASCAL VOC、CrowdHuman和 DOTA上分别将 AP分数提升了 +1.4、+0.6、+1.2，将 AP_{75} 提升了+1.5、+0.8、+1.0。需要留意，本文的方法在 PSC上没有显示加速。这可能是因为 DOTA数据集包含更多小物体，因此本文仅在 P4层级上实现 SlimHead。尽管如此，SlimHead在 PSC上仍实现了+1.2AP。

表 5.6: **SlimHead在3个检测数据集上的性能：**FPS是在单块 RTX 3090显卡上测量的。

数据集	SlimHead	AP	AP_{50}	AP_{75}	FPS
PASCAL VOC [36]		56.3	79.3	62.0	33.6
	✓	57.7	80.3	63.5	34.2
CrowdHuman [245]		44.0	78.8	43.3	33.6
	✓	44.6	79.1	44.1	34.2
DOTA [29]		41.9	68.2	42.9	31.3
	✓	43.1	68.8	43.9	30.1

SlimHead纳入实例分割

本文进一步将 SlimHead纳入实例分割方法中。这里使用 BoxInst [301]和 CondInst [302]两种实例分割模型，其具有 ResNet-50主干网络和 FPN颈部网络。本文遵循官方训练设置，采用 1× 训练计划。设置级别选择器 $K = 4$，并在检测头网络的所有层中使用 DCN [284]。结果由 COCO val2017上的 box AP和 mask AP报告。如表 5.7所示，SlimHead可以明显改善这两种实例分割模型的检测精度与分割精度。 BoxInst与 CondInst的 box AP分别被提高了0.9与1.6，而 mask AP则被分别被提高了0.7与1.2。这表明本文的 SlimHead具有良好的泛化能力，其不仅有利于目标检测任务，同时也有利于实例分割任务。更重要的是，SlimHead不会导致推理速度的下降，反而会略微加快模型效率。这再次证明了本文的方法在实践中具有高通用性与高效能。

表 5.7: **SlimHead在两种实例分割方法上的性能比较：**结果在 MS COCO val2017上报告。FPS是在单块 RTX 3090显卡上测量的。

实例分割模型	SlimHead	box			mask			FPS
		AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}	
BoxInst [301]		39.6	58.5	42.9	31.1	53.1	31.7	27.8
	✓	40.5	59.1	43.9	31.8	53.7	32.6	28.1
CondInst [302]		39.3	58.3	42.4	35.7	56.2	38.1	29.3
	✓	40.9	59.6	44.2	36.9	57.4	39.5	29.7

SlimHead纳入不同的主干网络

为了进一步验证适用性，本文将 SlimHead纳入到各种主干网络中。本文选择轻量级主干网络 ResNet-18 [108]、经典主干网络 ResNet-50、更重型

的 CNN主干网络 ResNext-101-64x4d [267]和 Transformer主干网络 Swin-L [307]。对于 ResNet-18和 ResNet-50，采用单尺度 1× 训练计划。对于 ResNext-101-64x4d和 Swin-L，采用多尺度 2× 训练计划。在本实验中，使用密集目标检测器 TOOD [131]为基础模型。在表 5.8中，本文报告了这4个主干网络上的检测结果以及 FPS。可以看出，SlimHead在4个主干网络上持续提高了检测精度，同时加快了推理速度。在 ResNet-18、ResNet-50、ResNext-101-64x4d和 Swin-L上，AP分别提高了 +1.1、+0.9、+1.0、+0.5。此外，SlimHead也带来了效率提升，FPS轻微提高。这再次表明，本文的方法可以在密集目标检测中实现更好的速度-精度权衡。

表 5.8: **SlimHead在4个不同主干网络上的显卡内存占用情况和检测性能：**结果在 MS COCO val2017上报告。FPS是在单块 RTX 3090显卡上测量的。**TS:** 训练计划设置。

主干网络	TS	SlimHead	显卡内存 (MB)	降低	AP	AP_{50}	AP_{75}	FPS
ResNet-18 [108]	1×		2,105		38.0	54.6	40.7	47.1
		✓	1,788	↓15.1%	39.1	55.4	42.4	47.4
ResNet-50 [108]	1×		3,967		42.3	59.6	45.9	33.6
		✓	3,653	↓8.0%	43.2	60.4	47.0	34.2
ResNeXt-101-64x4d [267]	2×		10,240		48.1	66.2	52.4	16.9
		✓	9,930	↓3.0%	49.1	67.1	53.5	17.3
Swin-L [307]	2×		6,518		50.1	68.8	54.6	6.6
		✓	6,198	↓4.9%	50.6	69.3	54.7	6.7

SlimHead节约显卡内存占用

如章节 5.3.4中所述，由于浅层级特征图分辨率降低，SlimHead还可带来更低的显卡内存占用量。 表 5.8报告了训练时期显卡内存占用量。这里每块显卡所处理的批量大小为2张图像。可以看到，本文的 SlimHead在 ResNet-18、ResNet-50、ResNeXt-101-64x4d、Swin-L上分别将显卡内存占用量减少了15.1%、8.0%、3.0%、4.9%。值得一提的是，随着模型变得愈发轻量级，显卡内存占用量的减少也变得愈发显著。这验证了本文方法的一个重要优势，即它可以节省显卡内存占用量，本文认为这对于低端边缘设备的部署有着重要价值。

与其它标签分配的比较

最后，本文将 SlimHead与各种标签分配算法进行比较。标签分配算法是所有当前最先进的目标检测器的基础。能否与标签分配算法共存使用，将是一个算法适用性的一个重要考量。为了公平比较，这里所有选定的目标检测器均采用 ResNet-50-FPN模型和单尺度 1× 训练计划。所有检测器均使用经典的密集目标检测器 FCOS [101]作为基础检测框架。 表 5.9报告了定量结果。可以看出，本文的 SlimHead的表现远远优于这些标签分配算法。值得注意的是，尽管由于目标检测器之间存在显著的结构差异，本文的方法不能直接扩展到基于查询的检测器当中，例如 DETR系列检测器 [117–120, 122–125, 142]，但表 5.9仍然展示了有希望的结果，即密集目标检测器在算法层面上仍然可以比基于查询的检测器表现更好。具体来说，DDQ-FCN [236]同样采用了 FCOS的结构，但却具有基于查询的学习范式。它遵循着与基于 DETR的检测器相同的一对一二分图匹配算法。 表 5.9的结果表明，如果检测网络在密集目标检测器和基于查询的检测器之间对齐，本文的 SlimHead（43.2 AP）仍然可以胜过基于查询的检测器，即 DDQ-FCN（41.5 AP）。

表 5.9: **与检测头网络中使用的各种标签分配算法进行定量比较：**结果在 MS COCO val2017上报告。FPS是在单块 RTX 3090显卡上测量的。

检测头网络	基础框架	标签分配算法	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	FPS
PAA [137]	FCOS	有锚框一对多	40.4	58.4	43.9	22.9	44.3	54.0	16.0
AutoAssign [290]	FCOS	有锚框一对多	40.4	59.6	43.7	22.7	44.1	52.9	34.5
OTA [138]	FCOS	有锚框一对多	40.7	58.4	44.3	23.2	45.0	53.6	36.9
DDQ-FCN [236]	FCOS	DETR 一对一	41.5	60.9	45.9	25.1	44.6	53.1	36.0
DW + box refine [132]	FCOS	有锚框一对多	42.1	59.9	45.1	24.2	45.3	55.9	35.0
TOOD [131]	FCOS	无锚框一对多	42.3	59.6	45.9	25.8	45.6	54.9	33.6
DyHead [235]	FCOS	无锚框一对多	42.6	60.1	46.4	26.1	46.8	56.0	24.3
SlimHead (Ours)	FCOS	无锚框一对多	**43.2**	60.4	47.0	24.2	47.0	58.5	34.2

5.6 本章小结

在本章中，本文重新审视了密集目标检测中流行的多层级学习框架。由于浅层级非常耗时，本章目标是探索检测头网络的本质属性，以及如何充分利用它们来实现更好的速度-精度权衡。通过一系列启发式的性能敏感性分析，本章提出了 SlimHead，这是一个非常简单、高效且易于推广的检测头网络，它进一

步释放了密集目标检测器多层级学习的潜力。 SlimHead有4大优势：1）降低多层级学习的计算复杂度；2）灵活结合改进的卷积算子操作以提高检测精度，例如 DCN；3）节省显卡内存占用量；4）高度可推广到各种检测器和数据集。在 PASCAL VOC、MS COCO、CrowdHuman、DOTA、通用/有向目标检测器和实例分割上进行的大量实验证明了本章所提方法的广泛适用性和很高的实用价值，从而为迈向高效能目标检测提供了一条切实可行的改进道路。

第六章 总结和展望

6.1 工作总结

目标检测作为经典老牌的计算机视觉任务，是现代业界所提倡的智能体的奠基石。目标检测作为机器理解现实世界的接口，本质上在于读取数据并归纳总结，它完成的是光学信号到语义信息的转化。它的拓展任务之纷繁，下游应用之丰富，部署场景之多样，使得它已深入渗透到了人们的日常生活之中。高效能，是目标检测发展的必由之路。一个强大的目标检测器必然经历着从低精度到高精度、从低效率到高效率的发展过程。对目标检测基础技术的研究也将意义非凡，通常蕴含着放射性的重大影响，有利于其拓展任务、下游应用、部署场景更好地展开与效能提升。

本文的研究为目标检测领域提供了新的视角，同时也为目标检测领域所面临的问题与挑战进行了一次深层次的分析。本论文对影响目标检测取得高效能的两大问题进行了探索，包括空间均衡与紧致表达，这其中涉及到影响检测鲁棒性的空间偏差、目标检测知识蒸馏的低效性、以及多层级学习的低效能问题。从技术的层面上共设计了五种方法，包括定位蒸馏、选择性区域蒸馏法、SlimHead检测头网络、区域评估、空间均衡学习。从分析的层面上共进行了五种探索，包括定位蒸馏与分类蒸馏的理论联系，logit模仿与特征模仿的优劣性比较，多层级学习检测头网络的敏感性分析，空间偏差的相关因素探索，以及空间失衡问题的建立。本论文的各部分贡献总结如下：

- 本文第二章主要介绍了目标检测的基础模型架构，以及学界在实现高效能目标检测方面所做出的一些努力，包括鲁棒性研究、边界框的表示与优化、知识蒸馏、以及多层级学习等。第二章还介绍了目标检测的常用数据集与评估方法，它们是构成目标检测研究的基础。
- 本文第三章介绍了一种新的区域评估方法，它是对传统平均精度指标的一种推广。基于区域评估，第四章揭示了目标检测取得高效能的新阻碍——空间偏差，并对其成因、幅度、来源进行了系统性的分析。随后，第四章为目标检测领域建立了一个新的研究课题——空间失衡问题。作为对该问

题的首个解决手段，第四章随后给出了空间均衡学习，以提高目标检测器的空间均衡性。

- 本文第四章介绍了用于提高目标检测知识蒸馏效能的定位蒸馏技术与选择性区域蒸馏法。随后，理论分析揭示了本文的定位蒸馏与分类蒸馏存在密切的联系，实验结果则展示了这种 logit模仿技术可超越以往强大的特征模仿技术。在优劣性分析方面，第三章还展示了 logit模仿与特征模仿在优化方面的差异性，为进一步研究目标检测知识蒸馏提供了新的见解与方案。
- 本文第五章介绍了影响密集目标检测器效能的一大阻碍在于多层级学习的低效能。通过一系列对检测头网络的性能敏感性分析，第五章总结了检测头网络的本质属性，并给出了一种简单优雅的方法，称为 SlimHead，来重新找到速度-精度的新权衡。它具有简单、高效、易于推广、节省显卡内存占用等优点。

6.2 研究展望

目标检测在当下依然是火热的研究领域，发展迅速。高效能目标检测将依旧是学界需要长期努力的研究课题。本文就目标检测空间偏差、知识蒸馏、多层级学习三大方面尝试改善目标检测的效能，以期望得到一个更加空间均衡与紧致表达的高效能目标检测器。尽管如此，目标检测依然存在许多挑战，仍亟待解决，本论文认为在未来值得进一步研究的的方向包括：

- 目标检测的空间偏差揭示了模型自身的鲁棒性缺陷。图像是二维的，而对于视频目标检测、目标跟踪等任务，是否存在三维的空间偏差？例如时间偏差，亦或者二者之综合的时空偏差。目前学界仍缺乏对此类鲁棒性问题的探索，那么也就相应地缺失对该问题的解决方案。
- 本文展示了定位蒸馏在密集目标检测上的可行性，而对于基于 DETR的检测器上的知识蒸馏同样值得研究。此外，本文还揭示了分类知识与定位知识分而治之、因地制宜地传递给学生模型将带来益处，而对于诸如实例分割等任务，是否存在专属于掩码的知识也应该被考虑进来同样也将值得研究。更进一步，选择性区域蒸馏法的思想还可以被推广至拥有 N 个任务的框架，每一个任务都可能拥有专属的知识需要分开蒸馏。本文相信任务导向的知识蒸馏将对未来视觉模型，特别是多任务集于一身的智能体的效能优化方面具有重要意义。

- 目标检测的多层级学习低效能问题普遍存在于采用此种学习范式的模型，如3D目标检测、目标跟踪等领域，未来对 SlimHead的进一步推广也将有很高的应用价值。

参考文献

[1] 杨学, 严骏驰, 基于特征对齐和高斯表征的视觉有向目标检测, 中国科学:信息科学 53(11) 2023, 2250–2265.

[2] 周治国, 马文浩, 一种多层多模态融合3d目标检测方法, 电子学报 52(3) 2024, 696–708.

[3] 尹宏鹏, 陈波, 柴毅, 刘兆栋, 基于视觉的目标检测与跟踪综述, 自动化学报 42(10) 2016, 1466–1489.

[4] 熊珍凯, 程晓强, 吴幼冬, 左志强, 刘家胜, 基于激光雷达的无人驾驶 3d 多目标跟踪, 自动化学报 49(10) 2023, 2073–2083.

[5] 陈卫东, 郭蔚然, 刘宏炜, 朱奇光, 基于改进 Mask R-CNN 的模糊图像实例分割的研究, 电子与信息学报 42(11) 2020, 2805–2812.

[6] H. Maeda, Y. Sekimoto, T. Seto, T. Kashiyama, H. Omata, Road damage detection and classification using deep neural networks with smartphone images, Computer-Aided Civil and Infrastructure Engineering 33(12) 2018, 1127–1141.

[7] T. Zeng, S. Li, Q. Song, F. Zhong, X. Wei, Lightweight tomato real-time detection method based on improved yolo and mobile deployment, Computers and Electronics in Agriculture 205 2023, 107625.

[8] Z. Zhou, Z. Song, L. Fu, F. Gao, R. Li, Y. Cui, Real-time kiwifruit detection in orchard using deep learning on android smartphones for yield estimation, Computers and Electronics in Agriculture 179 2020, 105856.

[9] A. Gupta, A. Anpalagan, L. Guan, A. S. Khwaja, Deep learning for object detection and scene perception in self-driving cars: Survey, challenges, and open issues, Array 10 2021, 100057.

[10] D. Fernandes, A. Silva, R. Névoa, C. Simões, D. Gonzalez, M. Guevara, P. Novais, J. Monteiro, P. Melo-Pinto, Point-cloud based 3d object detection and classification methods for self-driving applications: A survey and taxonomy, Information Fusion 68 2021, 161–191.

[11] E. Arnold, O. Y. Al-Jarrah, M. Dianati, S. Fallah, D. Oxtoby, A. Mouzakitis,

A survey on 3d object detection methods for autonomous driving applications, IEEE Transactions on Intelligent Transportation Systems 20(10) 2019, 3782–3795.

[12] M. H. Tanveer, A. Kovarovics, C. Koduru, R. C. Voicu, C. Chun, G. Mahdi, An analysis of robotic dog’s machine learning-based detection for pedestrians and vehicles, in: International Conference on Intelligent Computing and Robotics (ICICR), 2024.

[13] N. Rees, K. Thiyagarajan, S. Kodagoda, Robotic guide dog for real-time indoor object detection and classification with localization, in: IEEE Applied Sensing Conference (APSCON), 2024.

[14] C. Kyrkou, YOLOpeds: efficient real-time single-shot pedestrian detection for smart camera applications, IET Computer Vision 14(7) 2020, 417–425.

[15] H. H. Nguyen, T. N. Ta, N. C. Nguyen, H. M. Pham, D. M. Nguyen, et al., Yolo based real-time human detection for smart video surveillance at the edge, in: IEEE International Conference on Communications and Electronics (ICCE), 2021.

[16] R. Laroca, E. Severo, L. A. Zanlorensi, L. S. Oliveira, G. R. Gonçalves, W. R. Schwartz, D. Menotti, A robust real-time automatic license plate recognition based on the YOLO detector, in: International Joint Conference on Neural Networks (IJCNN), 2018.

[17] R. Laroca, L. A. Zanlorensi, G. R. Gonçalves, E. Todt, W. R. Schwartz, D. Menotti, An efficient and layout-independent automatic license plate recognition system based on the YOLO detector, IET Intelligent Transport Systems 15(4) 2021, 483–503.

[18] Y. Yuan, W. Zou, Y. Zhao, X. Wang, X. Hu, N. Komodakis, A robust and efficient approach to license plate detection, IEEE Transactions on Image Processing 26(3) 2016, 1102–1114.

[19] X. Ran, H. Chen, X. Zhu, Z. Liu, J. Chen, Deepdecision: A mobile deep learning framework for edge video analytics, in: IEEE INFOCOM 2018-IEEE conference on computer communications, 2018.

[20] J. Fang, Q. Liu, J. Li, A deployment scheme of yolov5 with inference optimiza-

tions based on the triton inference server, in: IEEE International Conference on cloud computing and big data analytics (ICCCBDA), 2021.

[21] A. Pansare, N. Sabu, H. Kushwaha, V. Srivastava, N. Thakur, K. Jamgaonkar, M. Z. Faiz, Drone detection using yolo and ssd a comparative study, in: International Conference on Signal and Information Processing (IConSIP), 2022.

[22] P. Zhu, L. Wen, D. Du, X. Bian, H. Fan, Q. Hu, H. Ling, Detection and tracking meet drones challenge, IEEE Transactions on Pattern Analysis and Machine Intelligence 44(11) 2021, 7380–7399.

[23] M. Bakirci, M. M. Ozer, Adapting swarm intelligence to a fixed wing unmanned combat aerial vehicle platform, in: Data Analytics and Computational Intelligence: Novel Models, Algorithms and Applications, Springer, 2023, pp. 433–479.

[24] A. Petrovski, M. Radovanović, A. Behlić, Application of drones with artificial intelligence for military purposes, in: International Scientific Conference od Defensive Technologies–OTEH, Vol. 2022, 2022.

[25] Y. Yun, L. Hou, Z. Feng, W. Jin, Y. Liu, H. Wang, R. He, W. Guo, B. Han, B. Qin, et al., A deep-learning-based system for indoor active cleaning, in: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2022.

[26] I. S. Singh, I. Wijegunawardana, S. B. P. Samarakoon, M. V. J. Muthugala, M. R. Elara, Vision-based dirt distribution mapping using deep learning, Scientific Reports 13(1) 2023, 12741.

[27] R. Mirjalili, M. Krawez, F. Walter, W. Burgard, Vlm-vac: Enhancing smart vacuums through vlm knowledge distillation and language-guided experience replay, arXiv preprint arXiv:2409.14096 (2024).

[28] M. Bakirci, I. Bayraktar, Boosting aircraft monitoring and security through ground surveillance optimization with yolov9, in: International Symposium on Digital Forensics and Security (ISDFS), 2024.

[29] G.-S. Xia, X. Bai, J. Ding, Z. Zhu, S. Belongie, J. Luo, M. Datcu, M. Pelillo, L. Zhang, DOTA: A large-scale dataset for object detection in aerial images, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

[30] A. E. Ibrahim, R. Shoitan, M. M. Moussa, H. A. Elnemr, Y. Im Cho, M. S. Abdallah, Object detection-based automatic waste segregation using robotic arm, International Journal of Advanced Computer Science and Applications 14(6) (2023).

[31] J. Yin, K. G. S. Apuroop, Y. K. Tamilselvam, R. E. Mohan, B. Ramalingam, A. V. Le, Table cleaning task by human support robot using deep learning technique, Sensors 20(6) 2020, 1698.

[32] C. Zhihong, Z. Hebin, W. Yanbo, L. Binyan, L. Yu, A vision-based robotic grasping system using deep learning for garbage sorting, in: Chinese control conference (CCC), 2017.

[33] K. J. Singh, D. S. Kapoor, M. Abouhawwash, J. F. Al-Amri, S. Mahajan, A. K. Pandit, Behavior of delivery robot in human-robot collaborative spaces during navigation., Intelligent Automation & Soft Computing 35(1) (2023).

[34] S. Protasov, P. Karpyshev, I. Kalinov, P. Kopanev, N. Mikhailovskiy, A. Sedunin, D. Tsetserukou, Cnn-based omnidirectional object detection for hermesbot autonomous delivery robot with preliminary frame classification, in: International Conference on Advanced Robotics (ICAR), 2021.

[35] Z. Li, B. Xu, D. Wu, K. Zhao, S. Chen, M. Lu, J. Cong, A yolo-ggcnn based grasping framework for mobile robots in unknown environments, Expert Systems with Applications 225 2023, 119993.

[36] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, A. Zisserman, The pascal visual object classes (voc) challenge, International Journal of Computer Vision 88(2) 2010, 303–338.

[37] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C. L. Zitnick, Microsoft COCO: Common objects in context, in: European conference on computer vision (ECCV), 2014.

[38] W. Fang, L. Ding, B. Zhong, P. E. Love, H. Luo, Automated detection of workers and heavy equipment on construction sites: A convolutional neural network approach, Advanced Engineering Informatics 37 2018, 139–149.

[39] F. Zhou, H. Zhao, Z. Nie, Safety helmet detection based on YOLOv5, in: IEEE International conference on power electronics, computer applications

(ICPECA), 2021.

[40] A. Hayat, F. Morgado-Dias, Deep learning-based automatic safety helmet detection system for construction safety, Applied Sciences 12(16) 2022, 8268.

[41] Y. Li, H. Wei, Z. Han, J. Huang, W. Wang, Deep learning-based safety helmet detection in engineering management based on convolutional neural networks, Advances in Civil Engineering 2020(1) 2020, 9703560.

[42] N. D. Nath, A. H. Behzadan, S. G. Paal, Deep learning for site safety: Real-time detection of personal protective equipment, Automation in construction 112 2020, 103085.

[43] X. Fan, M. Jiang, RetinaFaceMask: A single stage face mask detector for assisting control of the covid-19 pandemic, in: IEEE International Conference on Systems, Man, and Cybernetics (SMC), 2021.

[44] P. Nagrath, R. Jain, A. Madan, R. Arora, P. Kataria, J. Hemanth, SSDMNV2: A real time dnn-based face mask detection system using single shot multibox detector and mobilenetv2, Sustainable cities and society 66 2021, 102692.

[45] S. Singh, U. Ahuja, M. Kumar, K. Kumar, M. Sachdeva, Face mask detection using yolov3 and faster r-cnn models: Covid-19 environment, Multimedia Tools and Applications 80 2021, 19753–19768.

[46] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, C. I. Sánchez, A survey on deep learning in medical image analysis, Medical Image Analysis 42 2017, 60–88.

[47] Y. Liu, Z. Ma, X. Liu, S. Ma, K. Ren, Privacy-preserving object detection for medical images with faster r-cnn, IEEE Transactions on Information Forensics and Security 17 2019, 69–84.

[48] P. F. Jaeger, S. A. Kohl, S. Bickelhaupt, F. Isensee, T. A. Kuder, H.-P. Schlemmer, K. H. Maier-Hein, Retina u-net: Embarrassingly simple exploitation of segmentation supervision for medical object detection, in: Machine Learning for Health Workshop, 2020.

[49] R. Yang, Y. Yu, Artificial convolutional neural network in object detection and semantic segmentation for medical imaging analysis, Frontiers in Oncology 11 2021, 638182.

[50] J. Li, F. Tang, C. Zhu, S. He, S. Zhang, Y. Su, BP-YOLO: A real-time product detection and shopping behaviors recognition model for intelligent unmanned vending machine, IEEE Access (2024).

[51] S.-J. Horng, P.-S. Huang, Building unmanned store identification systems using yolov4 and siamese network, Applied Sciences 12(8) 2022, 3826.

[52] S. Yang, P. Luo, C.-C. Loy, X. Tang, Wider face: A face detection benchmark, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[53] P. Viola, M. J. Jones, Robust real-time face detection, InternationalJournal of Computer Vision (IJCV) 57 2004, 137–154.

[54] K. Zhang, Z. Zhang, Z. Li, Y. Qiao, Joint face detection and alignment using multitask cascaded convolutional networks, IEEE Signal Processing Letters 23(10) 2016, 1499–1503.

[55] H. Jiang, E. Learned-Miller, Face detection with the faster r-cnn, in: IEEE International Conference on Automatic Face & Gesture Recognition, 2017.

[56] P. Y. Ingle, Y.-G. Kim, Real-time abnormal object detection for video surveillance in smart cities, Sensors 22(10) 2022, 3862.

[57] K.-E. Ko, K.-B. Sim, Deep convolutional framework for abnormal behavior detection in a smart surveillance system, Engineering Applications of Artificial Intelligence 67 2018, 226–234.

[58] C. Huang, Z. Wu, J. Wen, Y. Xu, Q. Jiang, Y. Wang, Abnormal event detection using deep contrastive learning for intelligent video surveillance system, IEEE Transactions on Industrial Informatics 18(8) 2021, 5171–5179.

[59] J. Ma, W. Shao, H. Ye, L. Wang, H. Wang, Y. Zheng, X. Xue, Arbitrary-oriented scene text detection via rotation proposals, IEEE Transactions on Multimedia 20(11) 2018, 3111–3122.

[60] M. Liao, B. Shi, X. Bai, Textboxes++: A single-shot oriented scene text detector, IEEE Transactions on Image Processing 27(8) 2018, 3676–3690.

[61] W. Wang, E. Xie, X. Li, W. Hou, T. Lu, G. Yu, S. Shao, Shape robust text detection with progressive scale expansion network, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

[62] X. Zhou, C. Yao, H. Wen, Y. Wang, S. Zhou, W. He, J. Liang, East: an efficient and accurate scene text detector, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[63] J. Zhou, D. Gao, D. Zhang, Moving vehicle detection for automatic traffic monitoring, IEEE Transactions on Vehicular Technology 56(1) 2007, 51–59.

[64] Y. Tang, C. Zhang, R. Gu, P. Li, B. Yang, Vehicle detection and recognition for intelligent traffic surveillance system, Multimedia Tools and Applications 76 2017, 5817–5832.

[65] A. Mhalla, T. Chateau, S. Gazzah, N. E. B. Amara, An embedded computer-vision system for multi-object detection in traffic surveillance, IEEE Transactions on Intelligent Transportation Systems 20(11) 2018, 4006–4018.

[66] H. Song, H. Liang, H. Li, Z. Dai, X. Yun, Vision-based vehicle detection and counting system using deep learning in highway scenes, European Transport Research Review 11(1) 2019, 1–16.

[67] H. Li, P. Wang, C. Shen, Toward end-to-end car license plate detection and recognition with deep neural networks, IEEE Transactions on Intelligent Transportation Systems 20(3) 2018, 1126–1136.

[68] Z. Xu, W. Yang, A. Meng, N. Lu, H. Huang, C. Ying, L. Huang, Towards end-to-end license plate detection and recognition: A large dataset and baseline, in: European conference on computer vision (ECCV), 2018.

[69] S. M. Silva, C. R. Jung, License plate detection and recognition in unconstrained scenarios, in: European conference on computer vision (ECCV), 2018.

[70] P. Li, W. Zhao, Image fire detection algorithms based on convolutional neural networks, Case Studies in Thermal Engineering 19 2020, 100625.

[71] F. M. Talaat, H. ZainEldin, An improved fire detection approach based on yolo-v8 for smart cities, Neural Computing and Applications 35(28) 2023, 20939–20954.

[72] S. Wu, L. Zhang, Using popular object detection methods for real time forest fire detection, in: International Symposium on Computational Intelligence and Design (ISCID), Vol. 1, 2018.

[73] R. Xu, H. Lin, K. Lu, L. Cao, Y. Liu, A forest fire detection system based on

ensemble learning, Forests 12(2) 2021, 217.

[74] Y. Al-Smadi, M. Alauthman, A. Al-Qerem, A. Aldweesh, R. Quaddoura, F. Aburub, K. Mansour, T. Alhmiedat, Early wildfire smoke detection using different yolo models, Machines 11(2) 2023, 246.

[75] C. Bahhar, A. Ksibi, M. Ayadi, M. M. Jamjoom, Z. Ullah, B. O. Soufiene, H. Sakli, Wildfire and smoke detection using staged yolo model and ensemble cnn, Electronics 12(1) 2023, 228.

[76] D. Du, Y. Qi, H. Yu, Y. Yang, K. Duan, G. Li, W. Zhang, Q. Huang, Q. Tian, The unmanned aerial vehicle benchmark: Object detection and tracking, in: Proceedings of the European conference on computer vision (ECCV), 2018.

[77] X. Yang, J. Yang, J. Yan, Y. Zhang, T. Zhang, Z. Guo, X. Sun, K. Fu, Scrdet: Towards more robust detection for small, cluttered and rotated objects, in: IEEE/CVF International Conference on Computer Vision (ICCV), 2019.

[78] X. Yang, Q. Liu, J. Yan, A. Li, Z. Zhang, G. Yu, R3det: Refined single-stage detector with feature refinement for rotating object, in: AAAI conference on artificial intelligence (AAAI), 2021.

[79] X. Yang, J. Yan, Q. Ming, W. Wang, X. Zhang, Q. Tian, Rethinking rotated object detection with gaussian wasserstein distance loss, in: International Conference on Machine Learning (ICML), 2021.

[80] X. Yang, X. Yang, J. Yang, Q. Ming, W. Wang, Q. Tian, J. Yan, Learning high-precision bounding box for rotated object detection via kullback-leibler divergence, in: NeurIPS, 2021.

[81] D. Feng, C. Haase-Schütz, L. Rosenbaum, H. Hertlein, C. Glaeser, F. Timm, W. Wiesbeck, K. Dietmayer, Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges, IEEE Transactions on Intelligent Transportation Systems 22(3) 2020, 1341–1360.

[82] C. Badue, R. Guidolini, R. V. Carneiro, P. Azevedo, V. B. Cardoso, A. Forechi, L. Jesus, R. Berriel, T. M. Paixao, F. Mutz, et al., Self-driving cars: A survey, Expert systems with applications 165 2021, 113816.

[83] Z. Wang, Y. Kang, X. Zeng, Y. Wang, T. Zhang, X. Sun, Sar-aircraft-1.0: High-resolution sar aircraft detection and recognition dataset, Journal of Radars 12

2023, 906–922.

[84] S. Wei, X. Zeng, Q. Qu, M. Wang, H. Su, J. Shi, HRSID a high-resolution sar images dataset for ship detection and instance segmentation, IEEE Access 8 2020, 120234–120254.

[85] J. Li, C. Qu, J. Shao, Ship detection in sar images based on an improved faster r-cnn, in: SAR in Big Data Era: Models, Methods and Applications (BIGSARDATA), 2017.

[86] Y. Li, X. Li, W. Li, Q. Hou, L. Liu, M.-M. Cheng, J. Yang, Sardet-100k: Towards open-source benchmark and toolkit for large-scale sar object detection, in: NeurIPS, 2024.

[87] 蒋弘毅, 王永娟, 康锦煜, 目标检测模型及其优化方法综述, 自动化学报 47(6) 2021, 1232–1255.

[88] D. Lowe, Object recognition from local scale-invariant features, in: IEEE/CVF International Conference on Computer Vision (ICCV), 2001.

[89] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2005.

[90] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, D. Ramanan, Object detection with discriminatively trained part-based models, IEEE Transactions on Pattern Analysis and Machine Intelligence 32(9) 2010, 1627–1645. doi:10.1109/TPAMI.2009.167.

[91] A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, in: NeurIPS, 2012.

[92] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, Y. LeCun, Overfeat: Integrated recognition, localization and detection using convolutional networks, arXiv: 1312.6229 (2013).

[93] R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2014.

[94] J. R. Uijlings, K. E. Van De Sande, T. Gevers, A. W. Smeulders, Selective search for object recognition, International Journal of Computer Vision 104 2013, 154–

171.

[95] R. Girshick, Fast R-CNN, in: IEEE/CVF International Conference on Computer Vision (ICCV), 2015.

[96] S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: Towards real-time object detection with region proposal networks, NeurIPS 28 (2015).

[97] C. Yang, Z. Huang, N. Wang, Rapid object detection using a boosted cascade of simple features, in: Proceedings of the IEEE computer society conference on computer vision and pattern recognition, 2001.

[98] S. Belongie, J. Malik, J. Puzicha, Shape matching and object recognition using shape contexts, IEEE Transactions on Pattern Analysis and Machine Intelligence 24(4) 2002, 509–522. doi:10.1109/34.993558.

[99] H. Bay, T. Tuytelaars, L. Van Gool, Surf: speeded up robust features, in: ECCV, 2006.

[100] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2009.

[101] Z. Tian, C. Shen, H. Chen, T. He, FCOS: fully convolutional one-stage object detection, in: IEEE/CVF International Conference on Computer Vision (ICCV), 2019.

[102] Z. Cai, N. Vasconcelos, Cascade R-CNN: Delving into high quality object detection, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

[103] H. Law, J. Deng, CornerNet: Detecting objects as paired keypoints, in: European conference on computer vision (ECCV), 2018.

[104] J. Redmon, A. Farhadi, YOLOv3: An incremental improvement, arXiv: 1804.02767 (2018).

[105] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, S. Belongie, Feature pyramid networks for object detection, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[106] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, in: IEEE/CVF International Conference on Computer Vision (ICCV),

2017.

[107] J. Redmon, A. Farhadi, YOLO9000: better, faster, stronger, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[108] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[109] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: Unified, real-time object detection, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[110] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, A. C. Berg, SSD: Single shot multibox detector, in: European conference on computer vision (ECCV), 2016.

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

[112] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2015.

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

[114] S. Zhang, C. Chi, Y. Yao, Z. Lei, S. Z. Li, Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

[115] X. Li, W. Wang, L. Wu, S. Chen, X. Hu, J. Li, J. Tang, J. Yang, Generalized Focal Loss: learning qualified and distributed bounding boxes for dense object detection, in: NeurIPS, Vol. 33, 2020.

[116] A. Bochkovskiy, C.-Y. Wang, H.-Y. M. Liao, YOLOv4: Optimal speed and accuracy of object detection, arXiv: 2004.10934 (2020).

[117] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, S. Zagoruyko, End-to-end object detection with transformers, in: ECCV, 2020.

[118] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, J. Dai, Deformable DETR: Deformable

transformers for end-to-end object detection, in: International Conference on Learning Representations (ICLR), 2021.

[119] D. Meng, X. Chen, Z. Fan, G. Zeng, H. Li, Y. Yuan, L. Sun, J. Wang, Conditional DETR for fast training convergence, in: IEEE/CVF International Conference on Computer Vision (ICCV), 2021.

[120] Z. Dai, B. Cai, Y. Lin, J. Chen, Up-DETR: Unsupervised pre-training for object detection with transformers, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021.

[121] P. Sun, R. Zhang, Y. Jiang, T. Kong, C. Xu, W. Zhan, M. Tomizuka, L. Li, Z. Yuan, C. Wang, et al., Sparse R-CNN: End-to-end object detection with learnable proposals, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021.

[122] S. Liu, F. Li, H. Zhang, X. Yang, X. Qi, H. Su, J. Zhu, L. Zhang, DAB-DETR: Dynamic anchor boxes are better queries for detr, in: International Conference on Learning Representations (ICLR), 2022.

[123] F. Li, H. Zhang, S. Liu, J. Guo, L. M. Ni, L. Zhang, DN-DETR: Accelerate detr training by introducing query denoising, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022.

[124] H. Zhang, F. Li, S. Liu, L. Zhang, H. Su, J. Zhu, L. M. Ni, H.-Y. Shum, DINO: Detr with improved denoising anchor boxes for end-to-end object detection, in: International Conference on Learning Representations (ICLR), 2023.

[125] Z. Zong, G. Song, Y. Liu, DETRs with collaborative hybrid assignments training, in: ICCV, 2023.

[126] S. Chen, P. Sun, Y. Song, P. Luo, DiffusionDet: Diffusion model for object detection, in: IEEE/CVF International Conference on Computer Vision (ICCV), 2023.

[127] Y. Zhao, W. Lv, S. Xu, J. Wei, G. Wang, Q. Dang, Y. Liu, J. Chen, Detrs beat yolos on real-time object detection, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024.

[128] A. Wang, H. Chen, L. Liu, K. Chen, Z. Lin, J. Han, et al., YOLOv10: Real-time end-to-end object detection, NeurIPS 37 2024, 107984–108011.

[129] X. Li, W. Wang, X. Hu, J. Li, J. Tang, J. Yang, Generalized focal loss v2: Learning reliable localization quality estimation for dense object detection, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021.

[130] H. Zhang, Y. Wang, F. Dayoub, N. Sünderhauf, Varifocalnet: An iou-aware dense object detector, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021.

[131] C. Feng, Y. Zhong, Y. Gao, M. R. Scott, W. Huang, TOOD: Task-aligned one-stage object detection, in: IEEE/CVF International Conference on Computer Vision (ICCV), 2021.

[132] S. Li, C. He, R. Li, L. Zhang, A dual weighting label assignment scheme for object detection, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022.

[133] C.-Y. Wang, A. Bochkovskiy, H.-Y. M. Liao, YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023.

[134] C.-Y. Wang, I.-H. Yeh, H.-Y. Mark Liao, YOLOv9: Learning what you want to learn using programmable gradient information, in: European conference on computer vision (ECCV), 2024.

[135] X. Zhang, F. Wan, C. Liu, R. Ji, Q. Ye, Freeanchor: Learning to match anchors for visual object detection, in: NeurIPS, 2019.

[136] B. Zhu, J. Wang, Z. Jiang, F. Zong, S. Liu, Z. Li, J. Sun, Autoassign: Differentiable label assignment for dense object detection, arXiv preprint arXiv:2007.03496 (2020).

[137] K. Kim, H. S. Lee, Probabilistic anchor assignment with iou prediction for object detection, in: European conference on computer vision (ECCV), 2020.

[138] Z. Ge, S. Liu, Z. Li, O. Yoshie, J. Sun, OTA: Optimal transport assignment for object detection, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021.

[139] A. Vaswani, Attention is all you need, NeurIPS (2017).

[140] H. W. Kuhn, The hungarian method for the assignment problem, Naval research

logistics quarterly 2(1-2) 1955, 83–97.

[141] Q. Chen, X. Chen, J. Wang, S. Zhang, K. Yao, H. Feng, J. Han, E. Ding, G. Zeng, J. Wang, Group detr: Fast detr training with group-wise one-to-many assignment, in: IEEE/CVF International Conference on Computer Vision (ICCV), 2023.

[142] D. Jia, Y. Yuan, H. He, X. Wu, H. Yu, W. Lin, L. Sun, C. Zhang, H. Hu, DETRs with hybrid matching, in: CVPR, 2023.

[143] Y. Zhang, B. Kang, B. Hooi, S. Yan, J. Feng, Deep long-tailed learning: A survey, arXiv preprint arXiv:2110.04596 (2021).

[144] W. Ouyang, X. Wang, C. Zhang, X. Yang, Factors in finetuning deep model for object detection with long-tail distribution, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[145] Y. Li, T. Wang, B. Kang, S. Tang, C. Wang, J. Li, J. Feng, Overcoming classifier imbalance for long-tail object detection with balanced group softmax, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

[146] T. Wang, Y. Zhu, C. Zhao, W. Zeng, J. Wang, M. Tang, Adaptive class suppression loss for long-tail object detection, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021.

[147] B. Kang, S. Xie, M. Rohrbach, Z. Yan, A. Gordo, J. Feng, Y. Kalantidis, Decoupling representation and classifier for long-tailed recognition, in: International Conference on Learning Representations (ICLR), 2020.

[148] D. Mahajan, R. Girshick, V. Ramanathan, K. He, M. Paluri, Y. Li, A. Bharambe, L. Van Der Maaten, Exploring the limits of weakly supervised pretraining, in: European conference on computer vision (ECCV), 2018.

[149] Z.-H. Zhou, X.-Y. Liu, Training cost-sensitive neural networks with methods addressing the class imbalance problem, IEEE Transactions on Knowledge and Data Engineering 18(1) 2005, 63–77.

[150] Y. Cui, M. Jia, T.-Y. Lin, Y. Song, S. Belongie, Class-balanced loss based on effective number of samples, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

[151] A. Shrivastava, A. Gupta, R. Girshick, Training region-based object detectors with online hard example mining, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[152] J. Wang, K. Chen, S. Yang, C. C. Loy, D. Lin, Region proposal by guided anchoring, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

[153] J. Pang, K. Chen, J. Shi, H. Feng, W. Ouyang, D. Lin, Libra R-CNN: Towards balanced learning for object detection, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

[154] B. Li, Y. Liu, X. Wang, Gradient harmonized single-stage detector, in: AAAI conference on artificial intelligence (AAAI), 2019.

[155] Y. Cao, K. Chen, C. C. Loy, D. Lin, Prime sample attention in object detection, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

[156] A. Azulay, Y. Weiss, Why do deep convolutional networks generalize so poorly to small image transformations?, Journal of Machine Learning Research 20(184) 2019, 1–25.

[157] R. Zhang, Making convolutional networks shift-invariant again, in: International Conference on Machine Learning (ICML), 2019.

[158] M. A. Islam, M. Kowal, S. Jia, K. G. Derpanis, N. D. Bruce, Position, padding and predictions: A deeper look at position information in cnns, arXiv preprint arXiv:2101.12322 (2021).

[159] O. S. Kayhan, J. C. v. Gemert, On translation invariance in cnns: Convolutional layers can exploit absolute spatial location, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

[160] R. Xu, X. Wang, K. Chen, B. Zhou, C. C. Loy, Positional encoding as spatial inductive bias in gans, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021.

[161] R. G. Lopes, D. Yin, B. Poole, J. Gilmer, E. D. Cubuk, Improving robustness without sacrificing accuracy with patch gaussian augmentation, in: International Conference on Machine Learning Workshops (ICMLW), 2019.

[162] B. Alsallakh, N. Kokhlikyan, V. Miglani, J. Yuan, O. Reblitz-Richardson, Mind the pad–cnns can develop blind spots, in: International Conference on Learning Representations (ICLR), 2021.

[163] M. A. Islam, M. Kowal, S. Jia, K. G. Derpanis, N. D. Bruce, Global pooling, more than meets the eye: Position information is encoded channel-wise in cnns, in: IEEE/CVF International Conference on Computer Vision (ICCV), 2021.

[164] J. Choi, J. Lee, Y. Jeong, S. Yoon, Toward spatially unbiased generative models, in: IEEE/CVF International Conference on Computer Vision (ICCV), 2021.

[165] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, T. Aila, Analyzing and improving the image quality of stylegan, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

[166] T. Karras, S. Laine, T. Aila, A style-based generator architecture for generative adversarial networks, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

[167] G. Szabó, A. Horváth, Mitigating the bias of centered objects in common datasets, in: International Conference on Pattern Recognition (ICPR), 2022.

[168] M. Manfredi, Y. Wang, Shift equivariance in object detection, in: European conference on computer vision Workshops (ECCVW), 2020.

[169] Z. Wang, A. C. Bovik, H. R. Sheikh, E. P. Simoncelli, Image quality assessment: from error visibility to structural similarity, IEEE transactions on image processing 13(4) 2004, 600–612.

[170] H. R. Sheikh, A. C. Bovik, Image information and visual quality, IEEE Transactions on image processing 15(2) 2006, 430–444.

[171] R. Ferzli, L. J. Karam, A no-reference objective image sharpness metric based on the notion of just noticeable blur (jnb), IEEE transactions on image processing 18(4) 2009, 717–728.

[172] H. Liu, I. Heynderickx, Visual attention in objective image quality assessment: Based on eye-tracking data, IEEE transactions on Circuits and Systems for Video Technology 21(7) 2011, 971–982.

[173] W. Zhang, A. Borji, Z. Wang, P. Le Callet, H. Liu, The application of visual saliency models in objective image quality assessment: A statistical evalua-

tion, IEEE Transactions on Neural Networks and Learning Systems 27(6) 2015, 1266–1278.

[174] T.-J. Liu, K.-H. Liu, No-reference image quality assessment by wide-perceptual-domain scorer ensemble method, IEEE Transactions on Image Processing 27(3) 2017, 1138–1151.

[175] G. Zhai, X. Min, Perceptual image quality assessment: a survey, Science China Information Sciences 63 2020, 1–52.

[176] F. Lukas, Z. Budrikis, Picture quality prediction based on a visual model, IEEE Transactions on Communications 30(7) 1982, 1679–1692.

[177] C. Li, A. C. Bovik, Three-component weighted structural similarity index, in: Image quality and system performance VI, Vol. 7242, SPIE, 2009.

[178] E. C. Larson, D. M. Chandler, Most apparent distortion: full-reference image quality assessment and the role of strategy, Journal of electronic imaging 19(1) 2010, 011006–011006.

[179] A. Mittal, R. Soundararajan, A. C. Bovik, Making a “completely blind” image quality analyzer, IEEE Signal processing letters 20(3) 2012, 209–212.

[180] L. Chen, Z. Li, R. K. Maddox, Z. Duan, C. Xu, Lip movements generation at a glance, in: Eur. Conf. Comput. Vis., 2018.

[181] Y. Sun, A. Lu, L. Yu, Weighted-to-spherically-uniform quality evaluation for omnidirectional video, IEEE signal processing letters 24(9) 2017, 1408–1412.

[182] W. Xue, L. Zhang, X. Mou, A. C. Bovik, Gradient magnitude similarity deviation: A highly efficient perceptual image quality index, IEEE transactions on image processing 23(2) 2013, 684–695.

[183] D.-P. Fan, M.-M. Cheng, Y. Liu, T. Li, A. Borji, Structure-measure: A new way to evaluate foreground maps, in: Int. Conf. Comput. Vis., 2017.

[184] S. Bosse, D. Maniry, K.-R. Müller, T. Wiegand, W. Samek, Deep neural networks for no-reference and full-reference image quality assessment, IEEE Transactions on image processing 27(1) 2017, 206–219.

[185] X. Feng, T. Liu, D. Yang, Y. Wang, Saliency based objective quality assessment of decoded video affected by packet losses, in: 2008 15th IEEE International Conference on Image Processing, IEEE, 2008.

[186] K. Gu, S. Wang, H. Yang, W. Lin, G. Zhai, X. Yang, W. Zhang, Saliency-guided quality assessment of screen content images, IEEE Transactions on Multimedia 18(6) 2016, 1098–1110.

[187] W. Zhang, H. Liu, Toward a reliable collection of eye-tracking data for image quality research: Challenges, solutions, and applications, IEEE Transactions on Image Processing 26(5) 2017, 2424–2437.

[188] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, R. Timofte, Swinir: Image restoration using swin transformer, in: Int. Conf. Comput. Vis. Workshops, 2021.

[189] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, M.-H. Yang, Restormer: Efficient transformer for high-resolution image restoration, in: IEEE Conf. Comput. Vis. Pattern Recog., 2022.

[190] Y. Gu, X. Wang, L. Xie, C. Dong, G. Li, Y. Shan, M.-M. Cheng, VQFR: Blind face restoration with vector-quantized dictionary and parallel decoder, in: Eur. Conf. Comput. Vis., 2022.

[191] J.-X. Zhao, J.-J. Liu, D.-P. Fan, Y. Cao, J. Yang, M.-M. Cheng, Egnet: Edge guidance network for salient object detection, in: Int. Conf. Comput. Vis., 2019.

[192] Q. Hou, M.-M. Cheng, X. Hu, A. Borji, Z. Tu, P. H. Torr, Deeply supervised salient object detection with short connections, IEEE Transactions on Pattern Analysis and Machine Intelligence 41(4) 2019, 815–828.

[193] D.-P. Fan, G.-P. Ji, M.-M. Cheng, L. Shao, Concealed object detection, IEEE Transactions on Pattern Analysis and Machine Intelligence 44(10) 2022, 6024–6042.

[194] C. Dong, C. C. Loy, K. He, X. Tang, Image super-resolution using deep convolutional networks, IEEE Transactions on Pattern Analysis and Machine Intelligence 38(2) 2016, 295–307. doi:10.1109/TPAMI.2015.2439281.

[195] J. Kim, J. K. Lee, K. M. Lee, Accurate image super-resolution using very deep convolutional networks, in: IEEE Conf. Comput. Vis. Pattern Recog., 2016.

[196] Y. He, C. Zhu, J. Wang, M. Savvides, X. Zhang, Bounding box regression with uncertainty for accurate object detection, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

[197] J. Choi, D. Chun, H. Kim, H.-J. Lee, Gaussian YOLOv3: An accurate and

fast object detector using localization uncertainty for autonomous driving, in: IEEE/CVF International Conference on Computer Vision (ICCV), 2019.

[198] H. Qiu, H. Li, Q. Wu, H. Shi, Offset bin classification network for accurate object detection, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

[199] J. Yu, Y. Jiang, Z. Wang, Z. Cao, T. Huang, UnitBox: an advanced object detection network, in: ACM International Conference on Multimedia (ACM MM), 2016.

[200] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, S. Savarese, Generalized Intersection over Union: A metric and a loss for bounding box regression, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

[201] Z. Zheng, P. Wang, W. Liu, J. Li, R. Ye, D. Ren, Distance-IoU Loss: Faster and better learning for bounding box regression, in: AAAI conference on artificial intelligence (AAAI), 2020.

[202] B. Jiang, R. Luo, J. Mao, T. Xiao, Y. Jiang, Acquisition of localization confidence for accurate object detection, in: European conference on computer vision (ECCV), 2018.

[203] Z. Huang, L. Huang, Y. Gong, C. Huang, X. Wang, Mask scoring R-CNN, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

[204] E. Xie, P. Sun, X. Song, W. Wang, X. Liu, D. Liang, C. Shen, P. Luo, PolarMask: single shot instance segmentation with polar representation, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

[205] G. Hinton, O. Vinyals, J. Dean, Distilling the knowledge in a neural network, arXiv preprint arXiv:1503.02531 (2015).

[206] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, Y. Bengio, Fitnets: Hints for thin deep nets, in: International Conference on Learning Representations (ICLR), 2015.

[207] G. Chen, W. Choi, X. Yu, T. Han, M. Chandraker, Learning efficient object detection models with knowledge distillation, in: NeurIPS, Vol. 30, 2017.

[208] Q. Li, S. Jin, J. Yan, Mimicking very efficient network for object detection, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[209] T. Wang, L. Yuan, X. Zhang, J. Feng, Distilling object detectors with fine-grained feature imitation, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

[210] J. Guo, K. Han, Y. Wang, H. Wu, X. Chen, C. Xu, C. Xu, Distilling object detectors via decoupled features, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021.

[211] X. Dai, Z. Jiang, Z. Wu, Y. Bao, Z. Wang, S. Liu, E. Zhou, General instance distillation for object detection, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021.

[212] R. Sun, F. Tang, X. Zhang, H. Xiong, Q. Tian, Distilling object detectors with task adaptive regularization, arXiv preprint arXiv:2006.13108 (2020).

[213] D. Zhixing, R. Zhang, M. Chang, S. Liu, T. Chen, Y. Chen, et al., Distilling object detectors with feature richness, NeurIPS 34 2021, 5213–5224.

[214] G. Li, X. Li, Y. Wang, S. Zhang, Y. Wu, D. Liang, Knowledge distillation for object detection via rank mimicking and prediction-guided feature imitation, in: AAAI conference on artificial intelligence (AAAI), 2022.

[215] W. Park, D. Kim, Y. Lu, M. Cho, Relational knowledge distillation, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

[216] W. Cao, Y. Zhang, J. Gao, A. Cheng, K. Cheng, J. Cheng, PKD: General distillation framework for object detectors via pearson correlation coefficient, NeurIPS 35 2022, 15394–15406.

[217] G. Jocher, A. Chaurasia, A. Stoken, J. Borovec, Y. Kwon, K. Michael, J. Fang, C. Wong, Z. Yifu, D. Montes, et al., ultralytics/YOLOv5: v6. 2-yolov5 classification models, apple m1, reproducibility, clearml and deci. ai integrations, Zenodo (2022).

[218] P. Zhou, B. Ni, C. Geng, J. Hu, Y. Xu, Scale-transferrable object detection, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),

2018.

[219] G. Huang, Z. Liu, L. Van Der Maaten, K. Q. Weinberger, Densely connected convolutional networks, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[220] S. Liu, L. Qi, H. Qin, J. Shi, J. Jia, Path aggregation network for instance segmentation, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

[221] C.-Y. Fu, W. Liu, A. Ranga, A. Tyagi, A. C. Berg, DSSD: Deconvolutional single shot detector, arXiv:1701.06659 (2017).

[222] M. D. Zeiler, G. W. Taylor, R. Fergus, Adaptive deconvolutional networks for mid and high level feature learning, in: IEEE/CVF International Conference on Computer Vision (ICCV), 2011.

[223] G. Ghiasi, T.-Y. Lin, Q. V. Le, NAS-FPN: Learning scalable feature pyramid architecture for object detection, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

[224] D. H. Songtao Liu, Y. Wang, Learning spatial fusion for single-shot object detection, arxiv preprint arXiv:1911.09516 (2019).

[225] T. Kong, F. Sun, C. Tan, H. Liu, W. Huang, Deep feature pyramid reconfiguration for object detection, in: European conference on computer vision (ECCV), 2018.

[226] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

[227] M. Tan, R. Pang, Q. V. Le, Efficientdet: Scalable and efficient object detection, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

[228] M. Tan, Q. Le, EfficientNet: Rethinking model scaling for convolutional neural networks, in: International Conference on Machine Learning (ICML), 2019.

[229] T. Wang, X. Zhang, J. Sun, Implicit feature pyramid network for object detection, arXiv preprint arXiv:2012.13563 (2020).

[230] S. Bai, J. Z. Kolter, V. Koltun, Deep equilibrium models, NeurIPS 32 (2019).

[231] P. Viola, M. Jones, Querydet: Cascaded sparse query for accelerating high-resolution small object detection, in: IEEE/CVF Conference on Computer Vi-

sion and Pattern Recognition (CVPR), 2022.

[232] Q. Chen, Y. Wang, T. Yang, X. Zhang, J. Cheng, J. Sun, You only look one-level feature, in: CVPR, 2021.

[233] K. He, G. Gkioxari, P. Dollar, R. Girshick, Mask R-CNN, in: IEEE/CVF International Conference on Computer Vision (ICCV), 2017.

[234] Y. Wu, Y. Chen, L. Yuan, Z. Liu, L. Wang, H. Li, Y. Fu, Rethinking classification and localization for object detection, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

[235] X. Dai, Y. Chen, B. Xiao, D. Chen, M. Liu, L. Yuan, L. Zhang, Dynamic head: Unifying object detection heads with attentions, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021.

[236] S. Zhang, X. Wang, J. Wang, J. Pang, C. Lyu, W. Zhang, P. Luo, K. Chen, Dense distinct query for end-to-end object detection, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023.

[237] X. Zhang, F. Wan, C. Liu, R. Ji, Q. Ye, Freeanchor: Learning to match anchors for visual object detection, in: NeurIPS, Vol. 32, 2019.

[238] Z. Zheng, Y. Chen, Q. Hou, X. Li, P. Wang, M.-M. Cheng, Zone evaluation: Revealing spatial bias in object detection, IEEE Transactions on Pattern Analysis and Machine Intelligence (2024). doi:10.1109/TPAMI.2024.3409416.

[239] Z. Zheng, P. Wang, D. Ren, W. Liu, R. Ye, Q. Hu, W. Zuo, Enhancing geometric factors in model learning and inference for object detection and instance segmentation, IEEE Transactions on Cybernetics 52(8) 2022, 8574–8586.

[240] J. He, S. Erfani, X. Ma, J. Bailey, Y. Chi, X.-S. Hua, α-iou: A family of power intersection over union losses for bounding box regression, NeurIPS 34 2021, 20230–20242.

[241] Z. Zheng, R. Ye, P. Wang, D. Ren, W. Zuo, Q. Hou, M.-M. Cheng, Localization distillation for dense object detection, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022.

[242] Z. Zheng, R. Ye, Q. Hou, D. Ren, P. Wang, W. Zuo, M.-M. Cheng, Localization distillation for object detection, IEEE Transactions on Pattern Analysis and Machine Intelligence 45(8) 2023, 10070–10083. doi:10.1109/TPAMI.2023.

3248583.

[243] Z. Yang, Z. Li, X. Jiang, Y. Gong, Z. Yuan, D. Zhao, C. Yuan, Focal and global knowledge distillation for detectors, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022.

[244] J. Wang, Y. Chen, Z. Zheng, X. Li, M.-M. Cheng, Q. Hou, CrossKD: Cross-head knowledge distillation for dense object detection, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024.

[245] S. Shao, Z. Zhao, B. Li, T. Xiao, G. Yu, X. Zhang, J. Sun, Crowdhuman: A benchmark for detecting human in a crowd, arXiv preprint arXiv:1805.00123 (2018).

[246] A. Abdulkader, https://www.kaggle.com/datasets/parot99/face-mask-detection-yolo-darknet-format.

[247] Eunpyohong, https://www.kaggle.com/datasets/eunpyohong/fruit-object-detection.

[248] Alexander, https://www.kaggle.com/datasets/vodan37/yolo-helmethead/metadata.

[249] L. Huang, G. Liu, Y. Wang, H. Yuan, T. Chen, Fire detection in video surveillances using convolutional neural networks and wavelet transform, Engineering Applications of Artificial Intelligence 110 2022, 104737.

[250] S. Saponara, A. Elhanashi, A. Gagliardi, Real-time video fire/smoke detection based on cnn in antifire surveillance systems, Journal of Real-Time Image Processing 18 2021, 889–900.

[251] C. Badue, R. Guidolini, R. V. Carneiro, P. Azevedo, V. B. Cardoso, A. Forechi, L. Jesus, R. Berriel, T. M. Paixao, F. Mutz, et al., Self-driving cars: A survey, Expert Systems with Applications 165 2021, 113816.

[252] Y. Zhang, D. Zhou, S. Chen, S. Gao, Y. Ma, Single-image crowd counting via multi-column convolutional neural network, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[253] H. Idrees, M. Tayyab, K. Athrey, D. Zhang, S. Al-Maadeed, N. Rajpoot, M. Shah, Composition loss for counting, density map estimation and localization in dense crowds, in: European conference on computer vision (ECCV), 2018.

[254] J. Wan, Z. Liu, A. B. Chan, A generalized loss function for crowd counting and localization, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021.

[255] Q. Song, C. Wang, Z. Jiang, Y. Wang, Y. Tai, C. Wang, J. Li, F. Huang, Y. Wu, Rethinking counting and localization in crowds: A purely point-based framework, in: IEEE/CVF International Conference on Computer Vision (ICCV), 2021.

[256] M. T. Bhatti, M. G. Khan, M. Aslam, M. J. Fiaz, Weapon detection in real-time cctv videos using deep learning, IEEE Access 9 2021, 34366–34382.

[257] S. Narejo, B. Pandey, D. Esenarro Vargas, C. Rodriguez, M. R. Anjum, Weapon detection using yolo v3 for smart surveillance system, Mathematical Problems in Engineering 2021 2021, 1–9.

[258] L. Kirichenko, T. Radivilova, B. Sydorenko, S. Yakovlev, Detection of shoplifting on video using a hybrid network, Computation 10(11) 2022, 199.

[259] A. Chaman, I. Dokmanic, Truly shift-invariant convolutional neural networks, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021.

[260] A. Torralba, A. A. Efros, Unbiased look at dataset bias, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2011.

[261] K. Oksuz, B. C. Cam, S. Kalkan, E. Akbas, Imbalance problems in object detection: A review, IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 43(10) 2020, 3388–3415.

[262] A. Kuznetsova, H. Rom, N. Alldrin, J. Uijlings, I. Krasin, J. Pont-Tuset, S. Kamali, S. Popov, M. Malloci, A. Kolesnikov, et al., The open images dataset v4, InternationalJournal of Computer Vision (IJCV) 128(7) 2020, 1956–1981.

[263] S. Shao, Z. Li, T. Zhang, C. Peng, G. Yu, X. Zhang, J. Li, J. Sun, Objects365: A large-scale, high-quality dataset for object detection, in: IEEE/CVF International Conference on Computer Vision (ICCV), 2019.

[264] X. Yang, L. Hou, Y. Zhou, W. Wang, J. Yan, Dense label encoding for boundary discontinuity free rotation detection, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021.

[265] K. Chen, J. Wang, J. Pang, Y. Cao, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Xu, Z. Zhang, D. Cheng, C. Zhu, T. Cheng, Q. Zhao, B. Li, X. Lu, R. Zhu, Y. Wu, J. Dai, J. Wang, J. Shi, W. Ouyang, C. C. Loy, D. Lin, MMDetection: Open mmlab detection toolbox and benchmark, arXiv preprint arXiv:1906.07155 (2019).

[266] P. Goyal, P. Dollár, R. Girshick, P. Noordhuis, L. Wesolowski, A. Kyrola, A. Tulloch, Y. Jia, K. He, Accurate, large minibatch sgd: Training imagenet in 1 hour, arXiv preprint arXiv:1706.02677 (2017).

[267] S. Xie, R. Girshick, P. Dollár, Z. Tu, K. He, Aggregated residual transformations for deep neural networks, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[268] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, L. Shao, Pyramid Vision Transformer: A versatile backbone for dense prediction without convolutions, in: IEEE/CVF International Conference on Computer Vision (ICCV), 2021.

[269] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, S. Xie, A convnet for the 2020s, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022.

[270] D. Bau, B. Zhou, A. Khosla, A. Oliva, A. Torralba, Network dissection: Quantifying interpretability of deep visual representations, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[271] X. Cheng, Z. Rao, Y. Chen, Q. Zhang, Explaining knowledge distillation by quantifying the knowledge, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

[272] Q. Zhang, W. Wang, S.-C. Zhu, Examining cnn representations with respect to dataset bias, in: AAAI conference on artificial intelligence (AAAI), 2018.

[273] Z. Chen, C. Yang, Q. Li, F. Zhao, Z.-J. Zha, F. Wu, Disentangle your dense object detector, in: ACM International Conference on Multimedia (ACM MM), 2021.

[274] S. Zagoruyko, N. Komodakis, Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer, in: Interna-

tional Conference on Learning Representations (ICLR), 2017.

[275] J. Kim, S. Park, N. Kwak, Paraphrasing complex network: network compression via factor transfer, in: NeurIPS, 2018.

[276] X. Jin, B. Peng, Y. Wu, Y. Liu, J. Liu, D. Liang, J. Yan, X. Hu, Knowledge distillation via route constrained optimization, in: IEEE/CVF International Conference on Computer Vision (ICCV), 2019.

[277] G.-H. Wang, Y. Ge, J. Wu, Distilling knowledge by mimicking features, IEEE Transactions on Pattern Analysis and Machine Intelligence 44(11) 2021, 8183–8195.

[278] L. Zhang, K. Ma, Improve object detection with feature-based knowledge distillation: Towards accurate and efficient detectors, in: International Conference on Learning Representations (ICLR), 2020.

[279] Z. Kang, P. Zhang, X. Zhang, J. Sun, N. Zheng, Instance-conditional knowledge distillation for object detection, in: NeurIPS, 2021.

[280] Y. Zhou, X. Yang, G. Zhang, J. Wang, Y. Liu, L. Hou, X. Jiang, X. Liu, J. Yan, C. Lyu, W. Zhang, K. Chen, Mmrotate: A rotated object detection benchmark using pytorch, in: ACM International Conference on Multimedia (ACM MM), 2022.

[281] T. Furlanello, Z. Lipton, M. Tschannen, L. Itti, A. Anandkumar, Born again neural networks, in: International Conference on Machine Learning (ICML), 2018.

[282] L. Zhang, J. Song, A. Gao, J. Chen, C. Bao, K. Ma, Be your own teacher: Improve the performance of convolutional neural networks via self distillation, in: IEEE/CVF International Conference on Computer Vision (ICCV), 2019.

[283] H. Mobahi, M. Farajtabar, P. Bartlett, Self-distillation amplifies regularization in hilbert space, NeurIPS 2020, 3351–3361.

[284] X. Zhu, H. Hu, S. Lin, J. Dai, Deformable convnets v2: More deformable, better results, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

[285] S.-H. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, P. Torr, Res2net: A new multi-scale backbone architecture, IEEE Transactions on Pattern Anal-

ysis and Machine Intelligence (TPAMI) 43(2) 2021, 652–662. doi:10.1109/TPAMI.2019.2938758.

[286] C. Zhu, Y. He, M. Savvides, Feature selective anchor-free module for single-shot object detection, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

[287] C. Zhu, F. Chen, Z. Shen, M. Savvides, Soft anchor-point object detection, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

[288] J. Wang, W. Zhang, Y. Cao, K. Chen, J. Pang, T. Gong, J. Shi, C. C. Loy, D. Lin, Side-aware boundary localization for more precise object detection, in: European conference on computer vision (ECCV), 2020.

[289] H. Qiu, Y. Ma, Z. Li, S. Liu, J. Sun, Borderdet: Border feature for dense object detection, in: European conference on computer vision (ECCV), 2020.

[290] B. Zhu, J. Wang, Z. Jiang, F. Zong, S. Liu, Z. Li, J. Sun, Autoassign: Differentiable label assignment for dense object detection, arXiv preprint arXiv:2007.03496 (2020).

[291] K. Kim, H. S. Lee, Probabilistic anchor assignment with IoU prediction for object detection, in: European conference on computer vision (ECCV), 2020.

[292] J. Tang, R. Shivanna, Z. Zhao, D. Lin, A. Singh, E. H. Chi, S. Jain, Understanding and improving knowledge distillation, arXiv preprint arXiv:2002.03532 (2020).

[293] H. Li, Z. Xu, G. Taylor, C. Studer, T. Goldstein, Visualizing the loss landscape of neural nets, NeurIPS 31 (2018).

[294] S. I. Mirzadeh, M. Farajtabar, A. Li, N. Levine, A. Matsukawa, H. Ghasemzadeh, Improved knowledge distillation via teacher assistant, in: AAAI conference on artificial intelligence (AAAI), 2020.

[295] B. Zhao, Q. Cui, R. Song, Y. Qiu, J. Liang, Decoupled knowledge distillation, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022.

[296] Y.-H. Wu, Y. Liu, X. Zhan, M.-M. Cheng, P2t: Pyramid pooling transformer for scene understanding, IEEE Transactions on Pattern Analysis and Machine

Intelligence (TPAMI) 45(11) 2022, 12760–12771.

[297] Q. Hou, C.-Z. Lu, M.-M. Cheng, J. Feng, Conv2former: A simple transformer-style convnet for visual recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 46(12) 2024, 8274–8283.

[298] Z. Dai, H. Liu, Q. Le, M. Tan, Coatnet: Marrying convolution and attention for all data sizes, NeurIPS 2021, 3965–3977.

[299] M.-H. Guo, J.-X. Cai, Z.-N. Liu, T.-J. Mu, R. R. Martin, S.-M. Hu, Pct: Point cloud transformer, Computational Visual Media 7(2) 2021, 187 – 199.

[300] Y. Yu, F. Da, Phase-shifting coder: Predicting accurate orientation in oriented object detection, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023.

[301] Z. Tian, C. Shen, X. Wang, H. Chen, Boxinst: High-performance instance segmentation with box annotations, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021.

[302] Z. Tian, C. Shen, H. Chen, Conditional convolutions for instance segmentation, in: European conference on computer vision (ECCV), 2020.

[303] T. Kong, F. Sun, H. Liu, Y. Jiang, L. Li, J. Shi, Foveabox: Beyound anchor-based object detection, IEEE Transactions on Image Processing 29 2020, 7389–7398.

[304] M.-H. Guo, C.-Z. Lu, Z.-N. Liu, M.-M. Cheng, S.-M. Hu, Visual attention network, Computational Visual Media 9(4) 2023, 733–752.

[305] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, L. Shao, Pvt v2: Improved baselines with pyramid vision transformer, Computational Visual Media 8(3) 2022, 415–424.

[306] J. Wang, W. Zhang, Y. Cao, K. Chen, J. Pang, T. Gong, J. Shi, C. C. Loy, D. Lin, Side-aware boundary localization for more precise object detection, in: European conference on computer vision (ECCV), 2020.

[307] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: Hierarchical vision transformer using shifted windows, in: IEEE/CVF International Conference on Computer Vision (ICCV), 2021.

致谢

而立之年，博士毕业，感慨良多，挺值也挺爽的。我的家人们，以下是我由衷的感谢，一起共享我的欢乐吧。

首先，我要感谢我的导师程明明教授以及合作导师侯淇彬副教授，他们是我学术道路上的领航者。无论是论文的打磨，设备的帮助，奖项的支持，未来的指路，二位老师对我的科研、人生与职业发展道路都产生深远的影响。我还要感谢我的师弟陈宇铭，与他的合作甚是愉快，如有神助。我还要感谢天津大学数学学院的王萍教授，虽已从您那硕士毕业，但您对我的关怀一直延绵到了南开，谢谢。感谢天津大学智能与计算学部的任冬伟副教授，感谢您曾经对我学术能力的锻造。感谢上海交通大学的杨学，虽然我们隔着网线交流，但是你的帮助、经验与真知灼见让我学到不少东西。还要感谢已毕业的硕士师弟叶荣光，愿你事业安好。我还要感谢所有与我有过交流的同学们，你们的经验也让我获益良多。接下来，我要感谢我的父母，他们是我生活的力挺者，事业的支持者，无私的奉献者，永远爱你们。最后，稍微感谢一下自己吧。嘿嘿，我的运气全用在和你们相遇了。

有人说命运就像齿轮，而我要说命运更像铁索连环，每一环都是一位家人，一环套一环，环环相扣，缺一不可，是你们不断推着我向前走。

哎呀呀，不敢想象，博士就这么毕业了？我的学生时代，再见咯。

个人简历

郑兆晖，福建福州人，生于 1995 年 2 月 20 日。于 2018 年就读于天津大学数学学院，师从王萍教授，并于 2021 年获得数学硕士学位。于 2021 年进入南开大学攻读计算机科学与技术博士学位，师从程明明教授，目前主要研究方向为计算机视觉、目标检测。博士期间发表论文 5 篇，其中 CCF-A 类一作论文 3 篇，包括 2 篇 TPAMI 与 1 篇 CVPR。他担任 TPAMI、CVPR、ICCV、ICML、NeurIPS、ICLR、AAAI 等国际顶级期刊和会议审稿人。谷歌学术总引用 7100 余次，其中一作最高单篇引用 5000 余次，并入选 Paper Digest 最具影响力 AAAI 2020 论文 Top 2，另 1 篇 ESI 高被引论文，他引 1200 余次，并入选 Web of Science 对 IEEE TCYB 影响因子贡献项排行第一。他于博士期间获得 CCF-CV 学术新锐学者（全国 3 人）、博士生国家奖学金。他所开源的代码在社区累计获得星数超过 1600。

博士期间已发表学术论文：

[1] **Zhaohui Zheng***, Rongguang Ye*, Ping Wang, Dongwei Ren, Wangmeng Zuo, Qibin Hou, and Ming-Ming Cheng. Localization Distillation for Dense Object Detection. IEEE/CVF Conference on Computer Vision and Pattern Recognition (**CVPR**), 2022. (EI 源刊, CCF-A 类会议) 10.1109/CVPR52688.2022.00919

[2] **Zhaohui Zheng**, Rongguang Ye, Qibin Hou, Dongwei Ren, Ping Wang, Wangmeng Zuo, and Ming-Ming Cheng. Localization Distillation for Object Detection. IEEE Transactions on Pattern Analysis and Machine Intelligence (**IEEE TPAMI**), 2023. (SCI 一区，CCF-A 类期刊，影响因子 20.8) doi:10.1109/TPAMI.2023.3248583

[3] **Zhaohui Zheng**, Yuming Chen, Qibin Hou, Xiang Li, Ping Wang, and Ming-Ming Cheng. Zone Evaluation: Revealing Spatial Bias in Object Detection. IEEE Transactions on Pattern Analysis and Machine Intelligence (**IEEE TPAMI**), 2024. (SCI 一区，CCF-A 类期刊，影响因子 20.8) doi:10.1109/TPAMI.2024.3409416

[4] Yuxuan Li, Qibin Hou, **Zhaohui Zheng**, Ming-Ming Cheng, Jian Yang, and Xiang Li. Large Selective Kernel Network for Remote Sensing Object Detection. IEEE/CVF International Conference on Computer Vision (**ICCV**), 2023. (EI 源刊, CCF-A 类会议) 10.1109/ICCV51070.2023.01540

[5] Jiabao Wang, Yuming Chen, **Zhaohui Zheng**, Xiang Li, Ming-Ming Cheng, and Qibin Hou. CrossKD: Cross-Head Knowledge Distillation for Dense Object Detection. IEEE/CVF Conference on Computer Vision and Pattern Recognition (**CVPR**), 2024. (EI 源刊, CCF-A 类会议) 10.1109/CVPR52733.2024.01563

研究生期间主要获得的荣誉奖励：

1. CCF-CV 学术新锐学者（全国三人），2023
2. 博士生国家奖学金，2024