分类号：TP391　　单位代码：10363

密 级：公 开　　学　号:2210330106

题　目 基于多尺度特征增强的目标检测算法研究

题　目 Research on target detection algorithm based on multi-scale feature enhancement

学　生　姓　名：杨海燕

校　内　导　师：王凤随（教授）

校　外　导　师：徐文霞（工程师）

专　　业：电子信息

研　究　方　向：计算机视觉

论文答辩日期：2024 年 5 月 23 日

安徽工程大学学位论文原创性声明

本人郑重声明：我恪守学术道德，崇尚严谨学风。所呈交的学位论文，是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已明确注明和引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的作品及成果的内容。论文为本人亲自撰写，我对所写的内容负责，并完全意识到本声明的法律后果由本人承担。

学位论文作者签名：

日期：　　年　月　日

安徽工程大学学位论文版权使用授权书

学位论文作者完全了解学校有关保留、使用学位论文的规定，同意学校保留并向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅或借阅。本人授权安徽工程大学可以将本学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存和汇编本学位论文。

保密 □，在 ____ 年解密后适用本版权书。

本学位论文属于

不保密 □。

学位论文作者签名：　　指导教师签名：

日期：　　年　月　日　　日期：　　年　月　日

基于多尺度特征增强的目标检测算法研究

摘 要

目标检测作为一项利用计算机技术从图像数据中实现目标定位与识别的关键研究课题，旨在精确识别图像中各个对象的类别及其对应的空间位置信息。随着深度学习技术和卷积神经网络在图像处理领域的快速发展，使得目标检测技术获得了较高的识别准确度，同时推动了工业自动化、智能交通等领域的技术创新。然而，在实际应用中，由于目标的多样性、场景的复杂性以及目标形态的不规则性等因素的影响，使目标检测模型难以对不同尺度的目标信息进行有效地提取。为了充分提取和利用特征信息，论文主要采用多感受野、特征融合、特征金字塔以及注意力机制等方法增强对感兴趣目标特征的学习能力。在多类别检测任务中，从计算复杂度和性能两种不同的考量出发，选择计算成本低的无锚框网络 CenterNet 和性能高的一阶段检测网络 YOLOv5s 作为基线检测器，在小目标检测任务中，选择更能满足小目标对输出分辨率需求的 YOLOv5s 算法作为基线检测器，三者通过三种不同的多尺度特征增强模块提升模型的性能。其主要研究内容如下：

（1）针对 CenterNet 算法识别多尺度目标时缺乏有效感受野和对浅层位置信息的关注，使得模型表征能力不足的问题，提出了一种基于上下文增强和特征融合的改进方法。该方法采用多感受野和信息融合的思想，构建自适应上下文提取模块和特征融合策略。首先网络通过自适应上下文提取模块的多路径空洞卷积层获取目标的上下文特征，督促深层网络学习多尺度信息；然后，通过 ACON-C 激

活函数在网络中加入非线性因素，对网络神经元自适应地激活，增强网络的数据拟合能力；最后，采用联合注意力特征融合策略，旨在融合不同层次的特征信息，结合深层网络的语义信息与浅层网络的位置信息，以捕获对识别任务有用的多尺度特征信息。同时学习特征图在多个层次通道间的相关性，以加强网络对关键目标特征的专注度。所提方法在 PASCAL VOC 公开数据集上 mAP 达到 83.82%，约比基线算法 CenterNet 增加了 3.72%。与经典算法 Faster R-CNN、SSD、YOLOv3 相比 mAP 分别高出了 7.4%、9.5%、3.5%，优于现有的主要方法，充分验证了所提方法的有效性。

（2）针对 YOLOv5s 算法识别多尺度目标时缺乏对不同尺度特征间关联性的学习，导致信息冗余的问题，提出了一种基于多感受野注意力和特征融合的目标检测改进算法。该方法采用注意力机制、重参数和多尺度信息融合的思想，构建多感受野注意力模块和重新参数化的广义特征金字塔融合策略。首先网络通过多感受野注意力模块中并行的不同大小卷积核来获取高层网络的不同尺度特征，督促模型学习多尺度信息，并通过注意力结构来学习图像中不同尺度特征间的关联性，加强网络对有用信息的关注度，然后，联合重新参数化的广义特征金字塔融合网络对不同层次的特征信息进行合并，通过跳级连接和重参数化的方式来挖掘不同尺度下的特征信息，实现高级语义信息和低级空间信息的充分交换。最后，所提方法在 PASCAL VOC 公开数据集上 mAP 达到 87.1%，约比基线算法 YOLOv5s 增加了 1.8%，具有良好的识别准确度。

（3）针对 YOLOv5s 算法在识别安全帽佩戴行为时存在小目标细节信息提取能力不足的问题，提出了一种融合多尺度特征的小目标检测改进方法。该方法采用信息融合和优化损失函数的思想，构

建小目标检测层和多种损失协同监督策略。首先，网络通过小目标检测层融合来自深层网络的多尺度特征和更浅层的低维特征，以实现对图像四种尺度特征的学习，同时输出一个高分辨率的检测头，以保留丰富的细节信息辅助小目标判别，然后，采用 K-means 聚类算法重新生成尺度更精细的锚框参数来提高先验锚框与目标的匹配度。最后，在网络原有的定位损失中引入NWD损失，进一步优化网络模型，从而提升网络对小目标的定位准确度。所提算法在 SHWD 安全帽佩戴检测数据集上的平均精度（mAP）达到了 95.5%，比基线算法 YOLOv5s 提升了 2.7%。相较于经典算法 Faster R-CNN、SSD、YOLOv3 分别增加了 7.4%、9.5%、3.5%。改进的 YOLOv5s 相较于其他安全帽佩戴识别算法具有更高的识别准确度，在安全帽佩戴检测应用中具有良好的实用性。

关键词：目标检测，特征增强，多尺度感受野，注意力机制

RESEARCH ON TARGET DETECTION ALGORITHM BASED ON MULTI-SCALE FEATURE ENHANCEMENT

ABSTRACT

Target detection is a key research topic that uses computer technology to achieve target positioning and recognition from image data. It aims to accurately identify the categories of various objects in the image and their corresponding spatial location information. With the rapid development of deep learning technology and convolutional neural network in the field of image processing, target detection technology has achieved high recognition accuracy, and promoted technological innovation in the fields of industrial automation and intelligent transportation. However, in practical applications, due to the influence of factors such as the diversity of targets, the complexity of scenes, and the irregularity of target shapes, it is difficult for the target detection model to effectively extract target information at different scales. In order to fully extract and utilize the feature information, this paper mainly uses multi-scale receptive field, feature fusion, feature pyramid and attention mechanism to enhance the learning ability of the target features of interest. In the multi-category detection task, from the two different considerations of computational complexity and performance, the anchor-free network CenterNet with low computational cost and the one-stage detection network YOLOv5s with high performance are selected as the baseline detector. In the small target detection task, the YOLOv5s algorithm that can better meet the output resolution requirements of small targets is selected as the baseline detector.

The three enhance the performance of the model through three different multi-scale feature enhancement modules. The main research contents are as follows:

(1) Aiming at the problem that the CenterNet algorithm lacks effective receptive field and attention to shallow position information when identifying multi-scale targets, an improved method based on context enhancement and feature fusion is proposed. This method uses the idea of multi-receptive field and information fusion to construct an adaptive context extraction module and feature fusion strategy. Firstly, the network obtains the contextual features of the target through the multi-path hole convolution layer of the adaptive context extraction module, and urges the deep network to learn multi-scale information. Then, nonlinear factors are added to the network through the ACON-C activation function to adaptively activate the network neurons and enhance the data fitting ability of the network. Finally, the joint attention feature fusion strategy is adopted to fuse different levels of feature information, combined with the semantic information of the deep network and the location information of the shallow network to capture the multi-scale feature information useful for the recognition task. At the same time, the correlation between feature maps at multiple levels of channels is learned to enhance the network 's focus on key target features. The proposed method achieves mAP of 83.82% on the PASCAL VOC public dataset, which is about 3.72% higher than the baseline algorithm CenterNet. Compared with the classical algorithms Faster R-CNN, SSD, and YOLOv3, mAP is 7.4%, 9.5%, and 3.5% higher, respectively, which is better than the existing main methods, fully verifying the effectiveness of the proposed method.

(2) Aiming at the problem that the YOLOv5s algorithm lacks the learning of the correlation between different scale features when

identifying multi-scale targets, which leads to information redundancy, an improved target detection algorithm based on multi-sensory field attention and feature fusion is proposed. This method uses the idea of attention mechanism, re-parameter and multi-scale information fusion to construct a multi-receptive field attention module and a re-parameterized generalized feature pyramid fusion strategy. Firstly, the network fuses the multi-scale features from the deep network and the lower-dimensional features from the shallower network through the small target detection layer to realize the learning of the four scale features of the image, and outputs a high-resolution detection head to retain rich detail information to assist small target discrimination. Then, the joint re-parameterized generalized feature pyramid fusion network merges feature information at different levels, and mines feature information at different scales through skip connection and re-parameterization to achieve full exchange of high-level semantic information and low-level spatial information. Finally, the proposed method achieves 87.1% mAP on the PASCAL VOC public dataset, which is about 1.8% higher than the baseline algorithm YOLOv5s, and has good recognition accuracy.

(3) Aiming at the problem that the YOLOv5s algorithm has insufficient ability to extract the details of small targets when identifying the wearing behavior of safety helmets, an improved method of small target detection based on multi-scale features is proposed. This method adopts the idea of information fusion and optimization of loss function to construct a small target detection layer and a variety of loss collaborative supervision strategies. Firstly, the network fuses the multi-scale features from the deep network and the shallower low-dimensional features through the small target detection layer, and outputs a high-resolution detection head to retain rich details to assist small target discrimination. Then, the K-

means clustering algorithm is used to regenerate the anchor frame parameters with finer scale to improve the matching degree between the prior anchor frame and the target. Finally, the NWD loss is introduced into the original positioning loss of the network to further optimize the network model, thereby improving the positioning accuracy of the network for small targets. The average accuracy (mAP) of the proposed algorithm on the SHWD helmet wearing detection dataset reaches 95.5%, which is 2.7% higher than the baseline algorithm YOLOv5s. Compared with the classical algorithm Faster R-CNN, SSD and YOLOv3, it increased by 7.4%, 9.5% and 3.5% respectively. The improved YOLOv5s has higher recognition accuracy than other helmet wearing recognition algorithms, and has good practicability in the application of helmet wearing detection.

KEY WORDS: target detection, feature enhancement, multi-scale receptive field, attention

目录

摘 要 I

ABSTRACT IV

目录 VIII

第 1 章 绪论 1

1.1 研究背景与意义 1

1.2 国内外研究现状 3

1.2.1 传统目标检测方法 4

1.2.2 基于深度学习的目标检测方法 5

1.3 论文的主要研究内容及架构安排 9

1.4 论文的组织结构 10

第 2 章 目标检测相关技术分析 12

2.1 引言 12

2.2 卷积神经网络的理论工作 12

2.2.1 卷积层 13

2.2.2 池化层 15

2.2.3 激活函数 15

2.2.4 全连接层 17

2.3 经典网络结构 18

2.3.1 CenterNet 模型 18

2.3.2 YOLOv5 模型 21

2.4 评价指标 23

2.5 本章小结 24

第 3 章 基于上下文增强和特征融合的目标检测算法 25

3.1 引言 25

3.2 算法流程 25

3.2.1 网络结构设计 25

3.2.2 ACON 激活函数的原理.....26
3.2.3 自适应上下文提取模块.....27
3.2.4 注意力特征融合模块.....28
3.3 实验设置与结果分析.....30
3.3.1 实验设置与数据集.....30
3.3.2 消融实验.....31
3.3.3 与其他方法进行比较.....33
3.3.4 可视化结果.....35
3.4 本章小结.....36
第 4 章 基于多感受野注意力和特征融合的目标检测算法 37
4.1 引言.....37
4.2 算法流程.....37
4.2.1 网络结构设计.....37
4.2.2 SK 注意力模块.....38
4.2.3 重新参数化的广义特征金字塔网络 RepGFPN.....40
4.3 实验结果.....41
4.3.1 实验设置.....41
4.3.2 消融实验.....42
4.3.3 与其他方法进行比较.....43
4.3.4 主观评价结果.....45
4.4 本章小结.....46
第 5 章 融合多尺度特征的小目标检测算法 47
5.1 引言.....47
5.2 算法流程.....48
5.2.1 网络结构设计.....48
5.2.2 小目标检测层 small_Layer.....49
5.2.3 K-means 聚类算法 50
5.2.4 损失函数的改进.....51

5.3 实验设置与结果分析.....52
5.3.1 实验设置和数据集.....52
5.3.2 消融实验.....53
5.3.3 与其他方法进行比较.....55
5.3.4 可视化结果.....56
5.4 本章小结.....57
第 6 章 总结与展望.....59
6.1 工作总结.....59
6.2 工作展望.....60
参考文献 61
致谢 69
攻读学位期间取得的学术成果情况 70

第 1 章 绪论

1.1 研究背景与意义

伴随数字化时代的到来，互联网技术的广泛应用以及移动计算设备的普及，极大促进了计算机视觉技术的快速进展，并使其成为当今科技创新领域的焦点之一。其中，目标检测[1-3]作为计算机视觉领域的一个重要分支，在图像识别与分析中占据了中心地位，并在目标跟踪、实例分割和物体辨识等多项关键计算视觉任务中扮演了基础且广泛的角色。计算机视觉研究的宗旨在于发掘与开发高效能的智能系统，这些系统能够取代人类执行复杂的目标识别与定位工作，以实现或超越人眼的处理能力。在当前信息量巨大的社会背景下，图像处理与目标检测的复杂度和挑战性均有所提升，它们已成为计算视觉研究的核心议题。目标检测技术不限于处理图像目标的分类问题，还包括对物体空间位置的精准识别与捕捉，即目标的分类+定位[4-6]。通过结合机器学习算法[7]和数字图像处理[8]来模拟人类视觉器官和大脑的学习能力，旨在从图像数据中自动识别出指定类别的目标，并准确地标定其边界。随着深度学习特别是卷积神经网络[9]（Convolutional Neural Networks, CNN）在模式识别方面取得重大突破，目标检测算法的发展也得到了极大推动。基于深度学习的方法逐渐替代了传统的手动特征提取技术，现代的深度学习模型显示出了自动学习和提取特征的强大能力，显著提高了目标检测技术的性能，尤其在准确性和泛化能力方面取得了突破性的进展。

伴随着计算机视觉技术的进步，深度学习驱动的目标检测算法已经渗透众多的实践应用领域。例如，如图 1-1 展示的场景中可见，目标检测算法在农业智能化[10]、智能交通[11]、智慧医疗[12]、工业安防[13]等不同领域得到了广泛的应用。

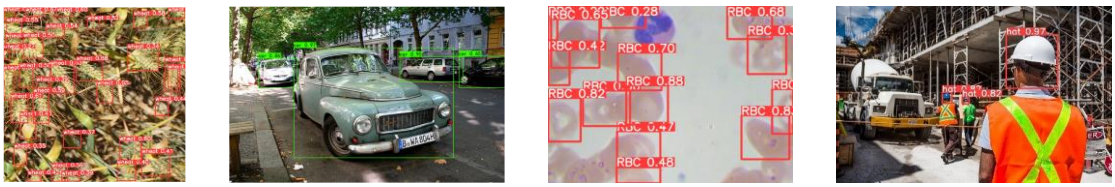

图 1-1 目标检测结果

（1）农业智能化。在迈向现代化的农业生产过程中，结合尖端设备与先进

计算技术的方法正变得日益普及，尤其是目标检测技术的应用，它在预测作物产量、检测农作物相关害虫、以及对这些害虫进行精确计数与分类方面，展现出巨大的潜力和实用价值。农林业不仅是我国基础产业的核心，更是推动国民经济持续发展的关键要素。害虫的周期性暴发对农作物的产量造成严重影响，同时也限制了农林产业的长远发展。广泛应用的化学农药虽然是防控害虫最有效的策略之一，但大量农业工作者对复杂多变的害虫种类缺乏足够认知，经常面临无法精确选用农药的问题。借助目标检测技术对害虫的精确监测，不仅可以提升农药使用的针对性和效率，还能在大幅降低经济成本的同时，显著减少对环境和生态系统可能造成的负面影响，为农业可持续发展注入了新动力。

（2）智能交通。在自动驾驶应用场景中，随着机动车辆数量的持续上升，解决交通拥堵问题、有效地指挥交通以及规范行人行为的挑战变得尤为迫切。在这一背景下，目标检测技术在自动驾驶系统中的环境感知部分扮演了极为关键的角色。这项前沿的技术具备了强大的识别能力，能够精准地定位道路上的关键要素，包括周围的车辆、行人、以及可能的障碍物，从而为交通流的智能化调控提供了可能。通过对目标检测结果进行深入的分析，自动驾驶车辆有效地避开了潜在的交通事故风险，并且在提升事故预警反应速度方面取得了显著的进展。更进一步，这项技术通过对各种紧急交通情况下的图像信息进行快速精确处理，极大提高了风险评估和信息处理的效率。目标检测技术的这些创新性功能在自动驾驶领域起到了至关重要的作用，不仅增强了行车的安全性，也显著提高了道路的交通效率。

（3）智慧医疗。在社会不断向前发展的同时，居民的生活方式也在逐步演变，愈加人们开始偏好于利用丰富的网络资源来轻松应对生活中的挑战，省时而减轻了日常劳动的负担。医疗行业也紧随这一变革趋势。尽管目前医学影像诊断仍重度依赖于医生丰富的临床经验，这一实践可能因医师之间经验水平的差异而波动诊断的准确率。而现代目标检测算法的涌现，它能够精确辨认出医学影像资料中的关键特征点，如诸多类型的肿瘤、病理性组织变化等，并在图像中标明它们确切的位置及体积大小。这种高级智能技术的运用，极大地增进了医疗界对各类疾病进行精确诊断的能力，有效降低了医护人员在检测流程中可能出现的疏漏以及误诊的风险。

（4）工业安防。目标检测技术在劳动安全监管这一关键领域扮演着一个至

关重要的角色。例如，在检查建筑工地或潜在危险场所工人是否正确戴上安全头盔时，该技术通过对摄像头捕获的实时视频流或图像进行深入剖析，能够以极高的准确性识别出场内工人的存在，并验证他们是否正确佩戴了安全防护装备。这样一个高效的机器视觉监测系统极大地优化了安全行为管理流程，提升了安全监管的效率和质量。利用高度精准的目标检测算法，系统能在第一时间准确定位到每一位工人的头部区域，然后进行进一步的精确分析，确认是否佩戴了保护头盔。一旦发现工人未按规定佩戴安全帽，系统将自动触发警报，并迅速将此信息告知监管责任人，以便他们能够立刻采取必要的应对措施，确保场地内劳动者的安全得到全面保护。这种创新性的智能监控技术不但显著提高了工作现场的安全管理水平，降低了因人为失误可能引发的各类安全隐患，也即时向安全管理人员提交了精准可靠的安全违规情况，大幅减少了工业生产中可能出现的事故风险，为保障工人的生命安全与健康做出了强有力的贡献。

随着目标检测技术的应用场景不断拓宽，人们对于检测模型的性能有了更高的期望。在众多目标检测的情形中，处理大小不一的物体变得尤为重要。以智能交通系统为例，在一个复杂的场景中，可能需要同时识别体型庞大的货车和较小的远处行人。在工业监控行业内，工人可能分布在镜头不同的深度范围内。虽然目前流行的目标检测算法在处理尺度差异小的物体时表现得出色，它们在面对具有显著尺度差异的物体时仍然显现出性能上的不足。这主要是因为传统检测模型依赖的是一个固定尺寸的感受野，往往难以精确地同时捕捉图像中存在的大幅度尺寸变化的目标，这种尺寸多样性直接影响了检测精度。因此，在目标检测领域中，对于不同尺度的物体进行有效检测，依旧是一个值得广泛探讨并充满挑战性的问题。

1.2 国内外研究现状

逾二十年的时间里，目标检测技术经历了从依赖手工设计特征的初期算法进化到采用深度学习框架的现代检测算法的转变。在此演变中，目标检测算法在提高识别精度与加快处理速度方面迈向了巨大的进步。多尺度目标检测算法因其固有的复杂性而成为研究领域的焦点。本节内容将细致回顾并分析传统目标检测方法和基于深度学习的目标检测方法的研究现状。这一过程不仅揭示了

目标检测技术的飞速发展，同时强调了对不同尺度目标检测的不断关注与研究。

1.2.1 传统目标检测方法

在目标检测技术发展的历程中，传统基于手工特征的方法曾经拥有中心地位，并且形成了一个标准的处理流程，其包含三个主要环节，如图 1-2 所呈现。首先，算法采取多尺度滑动窗口的策略，对整张图像进行系统性搜索，以期定位可能存在目标物体的初步区域。随后，利用专门设计的手工特征提取方法来处理这些候选区域。最后一步中，利用已经过训练的分类模型来对提取到的特征数据进行精确分类和标注。Viola-Jones 的人脸检测算法[14]便是基于这一方法学的优秀范例，其基于改良的 AdaBoost 算法[15]并通过调整滑动窗口大小以检测图像中的人脸。鉴于滑动窗口机制在计算上较为昂贵，Felzenszwalb 等人[16]提出了可变形部件模型（Deformable Parts Model，DPM），引入了一个更高效的策略：分解目标物体，并逐步推理出其不同组成部分。此外，Lowe 等人[17]开发的尺度不变特征变换（Scale Invariant Feature Transform，SIFT）算法，确保了图像在旋转或尺度变换后仍能够准确提取关键的局部信息。Lienhart 等人[18]提出的 Haar 特征检测器在实时人脸识别图（Histogram of Oriented Gradients，HOG）检测器[19]则通过调整图像大小来识别各尺度的标，尤其在行人检测领域展现了其越的性能。Cortes 等人[20]将 HOG 特征与支持向量机（Support Vector Machine，SVM）结合起来，这样的组合应用显著增强了对特定物体定位和识别的能力。

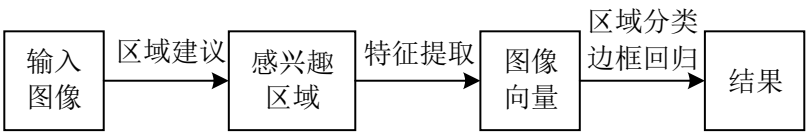

图 1-2 传统目标检测算法流程

虽然传统目标检测技术经过多年迭代已趋向成熟化，但其固有缺陷仍旧凸显。在进行潜在目标的区域选择环节时，传统的滑动窗口方法因其设计上的不够针对性而导致计算成本异常高昂，这不可避免地拖慢了整体的处理速度。而在特征抽取阶段，对于手工设计特征的过度依赖，造成了既耗时又劳神，并且由此设计出的特征往往在鲁棒性和泛化能力上有所欠缺，这限制了它们在多规模目标检测任务中的应用效果。面对这些难题，深度学习技术渐渐成为传统检测方法的有力替代者，并迅速崛起为目标检测领域的研究热点。借助于深度学

习的强大能力，研究者们获得了打破传统局限、显著提升检测效率与性能的可能性，尤以对复杂场景下的不同尺度目标检测任务，其表现更加卓越。

1.2.2 基于深度学习的目标检测方法

深度学习引领的目标检测技术已经稳固地成为现代计算机视觉领域的关键支柱，并且在近年来不断获得研究界的高度关注。深度学习的创新设计之处是赋予神经网络通过自我迭代训练来深入挖掘并理解数据间复杂的内部联系。这种策略显著胜过了以往依赖手工设计的初级特征提取技术，因为它能自动地从海量数据中识别出有效的特征表述，并不断优化学习机制，这一新进展在目标检测领域实现了重大飞跃。通过深度学习技术的应用，不仅实现了更快的检测速度，同时也显著提升了目标识别的准确性，这无疑是目标检测技术发展中的一大突破。卷积神经网络这一概念最初由 LeCun 等人[21]提出，但受制于当时计算力的限制，它未能在图像视觉领域得到即时广泛的认可。然而，随着 AlexNet[22]在一项规模庞大的图像识别比赛中取得了令人瞩目的成功，卷积神经网络便迅速在计算机视觉领域扩展开来。如图 1-3 所示，可以清晰地看出，2012 年成为了一道分水岭。在此之前，大多数检测算法仍然基于传统技术，但自 2012 年以来，深度学习算法开始主导这一领域。基于深度学习的目标检测算法可以根据是否需要预设锚框大致分为基于有锚框的(Anchor-Based)和基于无锚框的(Anchor-Free)目标检测算法；也可根据是否需要事先生成候选区域划分为两阶段（Two-Stage）和一阶段（One-Stage）目标检测算法。

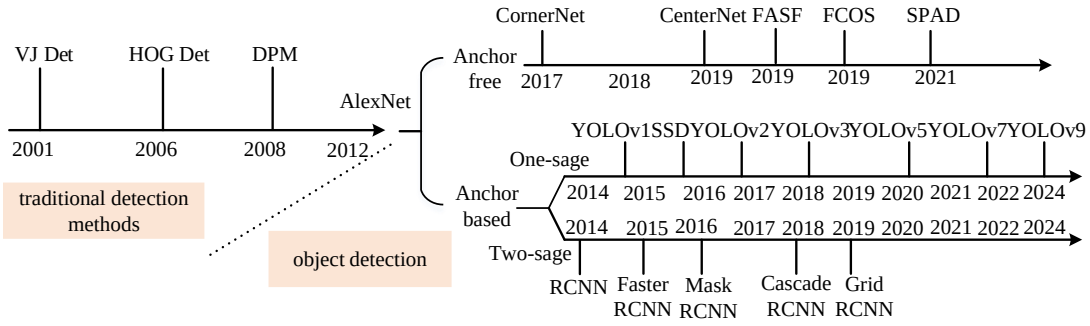

图 1-3 目标检测发展历程

（1）Anchor-Based 两阶段目标检测算法

随着深度学习领域的迅猛发展，基于锚框（Anchor-Based）的目标检测架构脱颖而出。这类先进的算法通过预先在待检测图像上设定一组多尺寸、多长

宽比例的锚框，将其作为初始的参考锚框。随后，针对每一个锚框进行类别识别以及位置微调，最终输出经过优化调整且极具代表性的锚框。根据检测操作流程的差异，基于锚框的检测算法又可以细分为两大类别：两阶段检测算法与一阶段检测算法。两阶段检测算法以其精确度高的特质闻名，在第一阶段产生可能涵盖目标的候选区域后，进而执行一系列精确的分类和位置校正操作，最终获得高度准确的检测成果。虽然其检测速度不及一阶段算法，但在精度要求较高的特定应用场景中表现出色。相较于两阶段检测算法，一阶段检测算法则通过省略候选区域提取的步骤，直接在预先设定的锚框基础上进行快速检测，以实现检测效率的显著提升，尽管在这一过程中可能会损失一部分精度。这两类算法依据其独特优势，为不同需求场景提供选择和应用的灵活性。

两阶段目标检测方法的核心在于一个分阶段的操作流程：首先，利用区域建议网络（Region Proposal Network，RPN）有效生成候选框，预测目标物体的大致位置。接着，进行分类与回归步骤，确定检测结果中目标的位置和类别。Girshick 等人[23]在 2014 年推出的 R-CNN 算法使两阶段检测概念取得里程碑式进展，它采取选择性搜索算法[24]（Selective Search，SS）从海量候选框中筛选潜在目标区域。经过统一缩放并通过深层卷积网络抽取特征后，采用分类器完成目标识别。R-CNN 在通用数据集中的表现相比传统技术有显著的提升，但它也暴露出一些明显的不足，如由选择性搜索产生的过多候选区域需要卷积网络对每一个独立提取特征，造成了大量计算上的冗余，模型训练也未能实现端到端，且检测速度缓慢。针对这种问题，He 等人[25]为了提升检测速度，引入了空间金字塔池化层（Spatial Pyramid Pooling Layer，SPP）来提升检测速度。空间金字塔池化通过有效减少所处理的候选区域数量，显著提高了处理速度，但端的训练问题仍需要进一步解决。随后，Fast-RCNN[26]通过 ROI 池化层和 Softmax 分类器的整合，实现了检测与回归器的联合训练，有效降低了计算代价。Ren 等人[27]提出 Faster-RCNN，通过集成 RPN 实现了自动生成区域建议，标志着实时目标检测领域的突破。尽管如此，由于 Faster-RCNN 对每个感兴趣区域进行独立计算，检测效率仍有进一步优化的空间。Lin 等人[28]设计的特征金字塔网络（FPN），通过自顶向下结构和横向连接，有效地将深层次的语义信息向浅层结构传播，进一步提高了检测精度。在 Faster-RCNN 基础上，研究者们又推出 Cascade-RCNN[29]、Mask-RCNN[30]、Libra-RCNN[31]等高效算法，不断推动检测

精度与速率的提升。尽管技术不断完善，两阶段检测算法在计算资源消耗上相对较大，速度在特定场合需求下依然有待提高。

（2）Anchor-Based 一阶段目标检测算法

在传统的两阶段目标检测方法之外，一阶段目标检测方法凭借其简约而高效的特色，省去了区域提案网络（Region Proposal Network，RPN）的中间步骤，依托于深度且精细的卷积网络架构，直接输出目标类别预测及其精确空间位置信息，实现了完整的端到端检测模式。OverFeat[32]作为一阶段检测器的先行者之一，巧妙处理了图像尺度的变换和集成化处理，灵活应对图像尺度的变化并有效进行目标检测。但 OverFeat 在分类器与回归器分离训练的方面，未能充分利用集成学习的潜力，局限了其综合性能。在 2016 年，Redmon 等人[33]提出了革命性的 YOLO 算法，这一算法将种类预测和坐标定位整合到一个的网络输出中，因其检测速度快而显著领先于传统的两阶段检测技术，尽管在小目标的精确性上略有不足，但在需要快速响应的场景中表现出色。随后，YOLOv2[34]通过采用更快的 Darknet-19 网络结构，并优化了分类精度与类别识别，进一步提升了表现。尽管性能有所增强，YOLOv2 在处理不同尺寸目标时显示出一定的性能缺陷。为追求更高速度与精准度，Liu 等人[35]于 2016 年提出 SSD 算法。该算法采用了多尺度特征图，并利用卷积进行分类与回归，从而显著提升了整体的检测精度。然而，SSD 在检测多尺度目标上的表现依旧未达到预期。在此环境下，Redmon 团队于 2018 年推出了 YOLOv3[36]，采用了 Darknet-53 网络，结合多尺度特征提取机制，增强了对不同尺寸目标的检测能力，扩大了在各行业的应用面。进入 2020 年，YOLOv4[37]和 YOLOv5[38]的发布，继续推进 YOLO 家族在检测精度和速度上的突破，2024 年 2 月，Wang 等[39]提出 YOLOv9，该算法利用信息瓶颈原理，通过实施可编程梯度信息来保留网络深度中重要数据，确保更可靠的梯度生成，提高了目标检测模型的收敛性和性能。以 SSD 和 YOLO 系列模型为代表的一阶段目标检测器，因其卓越的性能成为了业界关注的焦点，未来的发展将侧重于这两个方向的完善与创新。

（3）Anchor-Free 目标检测算法

锚框机制的提出有效提升了目标检测的性能，它通过预置一系列具有不同尺度和比例的锚框，精准地进行类别辨识和空间定位的微调。恰当的锚框参数能显著增强检测模型的能力，但这一配置过程依赖于先验知识，并在目标场景

变更时显示出局限性，需重新调整锚框设置，限制了模型的泛化性。此外，由于实际图像中目标有限，全局性设置锚框容易导致正负样本不均，成为影响算法训练效果的障碍。另外，样本划分的基准—交并比（Intersection over Union，IOU）的设定在训练与推理阶段亦会产生影响。为了更有效地解决这些问题，出现了无需预设锚框（Anchor-free）的目标检测算法。Anchor-free 指的是，无需为每个像素点设置不同尺寸的锚框，而是直接对像素点进行类别判断或匹配多个相同类别的像素点，从而避免了关于锚框的相关超参数设计过程。这种方法提高了模型对多类目标的适应性，并降低了模型的复杂性[40-41]。

在无锚框检测范畴中，Law 等人[42]提出了具有创新意义的 CornerNet 算法，开辟了该领域的研究之路。CornerNet 算法摒弃了传统的锚框预测方法，转而独立识别物体的左上角和右下角关键点，并使用嵌入式向量进行角点对匹配，预测物体的边界框坐标。该方案彻底放弃了锚框，将目标检测转换为关键点检测任务，为后续研究提供了新的视角。尽管 CornerNet 算法摆脱了锚框的依赖，但未考虑物体的关键内部特征，提高了误判风险，此外，它也在处理物体关键点被遮挡时显现出不足。为了克服这些问题，Zhou 等人[43]提出的 CenterNet 算法则关注于预测物体的中心点，利用深层特征，通过识别热力图上的局部最大值点简化处理流程，可适用于 2D 目标检测、3D 物体检测及人体姿势估计等任务。然而，CenterNet 的实操中存在下采样导致的中心点重合问题。为此，Zhu 等人[44]设计了 FASF 框架，采用自动化的特征层分配机制，动态选取最佳特征层，整合输出以达到最终检测效果。此外，Tian 等人[45]基于全卷积网络推出了 FCOS 检测器，采用逐像素点预测和锚点代替传统的锚框，避免了锚框相关的超参数设定。在实际的检测流程中，每个锚点同时承担分类和回归任务，并添加预测分支以评估锚点的中心得分，表示锚点到物体中心的距离。在推理阶段，将锚点预测的类别置信度与中心得分相乘得到一个数值，用作非极大值抑制排序的依据，以确保离目标中心更近的锚点识别出的物体边界框优先被选取和保留。深度学习的不断发展让目标检测性能获得重大提升，然而针对不同尺度物体的检测仍存在技术挑战。一阶段检测器虽然对尺度差异小的目标表现出色，面对一个图像内尺寸悬殊的物体时，由于感受野固定，传统的一阶段检测模型常常难以妥善处理，直接制约了其在全局性能上的表现。无锚框算法尽管采用关键点的方法动态预测物体宽高，但仅依靠单一尺度特征输出结构处理图像特

征，遂造成表征力不足，尤其是在保留浅层位置信息上存在损失。因此，如何进一步增强模型对各种尺度物体的适应性，持续提升多尺度物体的检测精度，便成了计算机视觉领域亟待攻克的一大挑战。

1.3 论文的主要研究内容及架构安排

不同尺度检测是目标检测领域较为活跃的主题之一，对不同尺度的目标进行检测，要求检测网络对尺度具有鲁棒性。因此不同尺度目标特征的学习是提升检测性能的关键。为此，本文提出了基于多尺度特征增强的目标检测算法研究，论文分别从 PASCAL VOC 数据集检测和安全帽佩戴检测两个场景出发，以 CenterNet 模型和 YOLOv5s 模型作为基线，从优化多尺度特征提取能力方面提升检测算法性能。论文探讨了目标检测算法精度优化研究过程中涉及的相关内容以及相关算法所能实现的功能。研究的主要内容概括如下：

（1）基于上下文增强和特征融合的目标检测算法。从充分利用感兴趣目标的上下文特征和减少浅层信息损失的角度出发，对 CenterNet 无锚框目标检测算法进行改进，提出一种基于上下文增强和特征融合的目标检测算法。首先，采用多感受野特征学习策略，通过在骨干网络 ResNet-50 后引入自适应上下文提取模块的多分支空洞卷积层来提取目标的上下文特征，并融合不同分支信息，以综合捕获多个尺度感受野下的上下文信息；其次，引入 ACON-C 激活函数引入非线性元素，以增强网络的数据拟合能力。最后，为减少浅层位置信息的损失，利用注意力特征融合策略将自适应上下文提取模块获取的多尺度上下文信息与不同深度的浅层特征进行融合，并学习不同维度特征图通道间的相关性来加强网络对关键信息的专注度，从而将多尺度特征图的有用信息进行合并。

（2）基于多感受野注意力和特征融合的目标检测算法。从弱化目标间的相关干扰和减少重要目标信息损失的角度出发，对 YOLOv5s 进行改进，提出一种基于多感受野注意力和特征融合的目标检测算法。通过采用注意力机制、重参数和多尺度信息融合的思想来优化网络，以达到有效提取特征以及最大化融合有用信息的目的。首先，考虑到不同感受野下的特征相互影响、特征权重分配不当会引起某一类目标特征的过度强调对其他类目标产生的干扰，将多感受野注意力模块嵌入到主干网络后，在学习多尺度信息的同时，通过注意力机制来

学习图像中不同尺度特征的关联性，给不同级别特征赋予适当的特征权重，以保证对多尺度特征的有效提取。其次，为减少重要目标信息的丢失，使用重新参数化的广义特征金字塔融合网络代替原来的颈部特征融合网络，通过跳级连接和重参数化的方式来挖掘上下层以及相邻层的特征信息，合并后的特征再传递到不同尺度的输出层，来保证多尺度特征信息的充分融合。最后，使用PASCAL VOC 数据集训练模型。

（3）融合多尺度特征的小目标检测算法。针对安全帽佩戴检测中小目标多的特点，从增强网络对小目标特征的提取能力和减少小目标位置偏差敏感度的角度出发，对 YOLOv5s 网络进行改进，提出了一种融合多尺度特征的小目标检测改进方法。首先，通过小目标检测层融合来自深层网络的多尺度特征和浅层网络的细节信息，以减少小目标信息损失。然后，为了提高锚框匹配度，采用 K-means 聚类算法重新生成尺度更精细的锚框参数。最后，使用归一化高斯距离损失度量真实框和预测框的距离，弱化小目标的位置偏差敏感度，再联合完全交并比损失提高定位准确度。

1.4 论文的组织结构

基于研究的主要内容，论文编排了六个章节，各章节内容概述如下：

第一章是绪论，首先介绍了目标检测研究的背景及重要性，接着简要概述了传统目标检测方法和基于深度学习的目标检测方法在国内外的研究现状，包括 Anchor-Based 两阶段目标检测器、Anchor-Based 一阶段目标检测器和 Anchor-Free 目标检测器的发展历史。最后，简洁概括了论文的主要研究内容。

第二章是对目标检测相关技术的介绍。首先介绍了卷积神经网络的技术原理，并阐述了其四个组成部分，其次，介绍了两种经典网络 CenterNet、YOLOv5 的网络架构，最后，介绍了评估目标检测算法性能的常用指标。

第三章着重研究了基于上下文增强和特征融合的目标检测算法，提出了自适应上下文提取和注意力特征融合的模块。该章节从算法流程和算法原理的角度进行了详细介绍，并分别从客观和主观角度对实验结果进行了分析。

第四章则基于多感受注意力和特征融合的目标检测算法研究。鉴于前一章算法输出固定分辨率特征的局限性，本章构建了对 YOLOv5s 网络进行改进的模

型。首先介绍了该方法的设计思路和算法流程，然后介绍了多感受野注意力和重新参数化广义特征金字塔的架构和原理，最后从消融实验、对比实验和可视化实验评估本章方法。

第五章是融合多尺度特征的小目标检测算法研究，受第三章、第四章算法采用特征融合思想的启发，本章考虑到融合更浅层的特征设计了小目标检测层。首先介绍了整体的算法框架和原理；其次阐述了小目标检测层的结构和损失函数的设计。最后在 SHWD 数据集上进行验证和分析。

第六章展望未来的工作，总结了本论文的研究内容和主要工作，并指出了当前研究的不足之处，并提出了今后的工作方向。

第 2 章 目标检测相关技术分析

2.1 引言

相较于传统的检测手段，深度学习以其强大的特征提取能力，在目标检测领域取得了重大进展。这一技术的引入，不仅极大地丰富了特征信息，更推动了目标检测技术的跨越式发展。本章主要剖析目标检测领域的核心理论，阐述卷积神经网络的基本原理，并介绍几种具有里程碑意义的经典网络结构。通过这一系统性的梳理与阐述，旨在为后续章节的研究工作提供坚实而深入的理论基础与指导。

2.2 卷积神经网络的理论工作

在神经科学研究的深刻启示下，卷积神经网络（CNN）的构建哲学深受人类大脑视觉系统高度有序与效能极佳的神经结构所影响。人脑的视觉信息处理核心，即视觉皮层，充斥着大量功能各异的神经元集群，它们按照视觉信息的解析需要，展现出层级的专业化和功能分化。在人的视觉观察过程中，多个神经元逐级激活，每层负责探测特定的视觉特性，例如形状、纹理、颜色及边缘等。这种分级加工的神经机制赋予了我们丰富而精确的视觉理解能力，并使得我们能深刻认知被观察物体的真实内质。故此，卷积神经网络于 20 世纪 90 年代初问世，迅速成为图像识别应用领域的重要成员。作为一种前馈型神经网络的变体，CNN 在图像处理领域内角色举足轻重，携带独立提取图像特征的能力，并展现了在图像分类上的优秀学习潜力。目前，CNN 的应用范畴已不仅限于图像解析，而是延伸到目标检测、自然语言处理等众多科学研究与工业应用场合。卷积神经网络的结构主要由若干层级协同工作的单元构成，包括输入数据的输入层，特征提取关键的卷积层，特征显著化的池化层，引入非线性以提高表达能力的激活层，整合信息的全连接层，以及结果输出的输出层。如图2-1所示，各层通过精密设计，赋予 CNN 类似于人类视觉系统的特征分解与智能解读能力，从而能够在复杂的计算任务中表现出色。

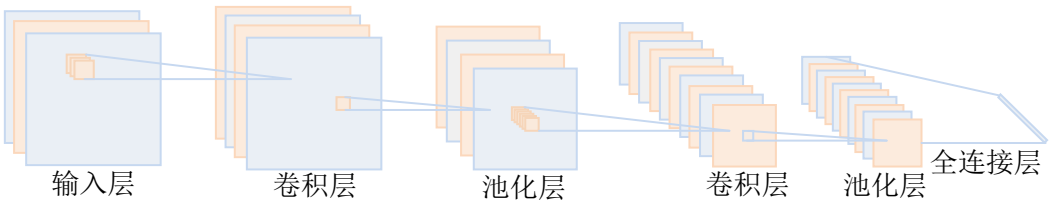

图 2-1 CNN 基础结构

2.2.1 卷积层

在构筑卷积神经网络（CNN）的多层结构中，卷积层充当着中枢的作用，不仅承载了网络大部分的计算任务，而且在信息的提炼处理方面发挥着关键作用。通过设计周全的卷积计算操作，卷积层能有效地从输入图像中抽取有关特征的维度信息，并生成一系列信息丰富的特征图谱集。在特征提取的初期阶段，初级特征图集中捕捉图像中的基础结构，如边缘、直线或曲线等元素；与此相对，随着层数的增加，深层特征图则专司从图像中提炼更高级的语义信息，这些信息对于理解图像的复杂空间构造至关必要，同时对图像的精确分类与辨识也具有决定性的影响。在信号处理领域，所指的卷积运算是一种特殊的数学操作，即在定义的操作域内，对两个函数实施一连串的乘积和加权求和，以获得最终的叠加效果，其详细节的运算步骤如数学表达式（2-1）、（2-2）所示。

$$f(x)*g(x)=\int_{-\infty}^{\infty} f(\tau)g(x-\tau)(x-\tau)d\tau \tag{2-1}$$

$$f(x)*g(x)=\sum_{\tau=-\infty}^{\infty} f(\tau)g(x-\tau)d\tau \tag{2-2}$$

其中，公式（2-1）表示卷积的连续运算，而公式（2-2）则是卷积的离散形式。在深度学习的大体框架之内，卷积运算作为核心算法，其基础理念在很大程度上源自于信号处理学科当中相类似的数学原理，然而在深度学习范畴中，卷积的实施和传统方法略有不同。在进行深度学习卷积时，应用的卷积核或滤波器不再需要遵循传统的信号处理中所需的翻转步骤。相反地，滤波器是以预设的跨度参数，在输入图像的辽阔数据矩阵上灵活地滑过，并进行搜索工作。滤波器在其每一个滑行步骤中，均会与其覆盖区域内的像素元素进行一一对应的乘积运算，并将这些离散的乘积值经过聚合，形成一个呈现新特性的特征图。具体的卷积运算过程如图 2-2 所示。

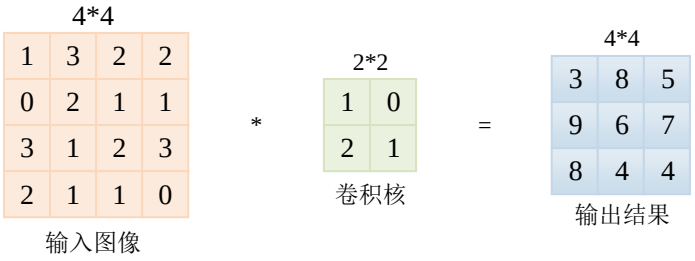

图 2-2 二维卷积计算过程

二维卷积的操作细节可通过如下公式（2-3）所示。

$$f(x,\ y)=(X*k)(x,y)=\sum_{i}\sum_{j}X(x-i,y-j)k(i,j) \tag{2-3}$$

其中，X 表示输入图像的具体维度参数；k 为选定的卷积核，它定义了过滤效果的范畴与力度；x 及 y 分别代表图像平面内，各个像素点所处的行列坐标索引；而 i 与 j 则精确到卷积核内部，指向卷积操作中的特定位置。基于输出特征图的预期尺寸要求，可以灵活设置卷积核的尺寸、移动步长及边界填充的额度，计算原理如公式（2-4）所示。

$$Y=\frac{X-k+2p}{d}+1 \tag{2-4}$$

在公式中，Y 代表在卷积过程中形成的输出特征图像；p 代表边缘填充值，它的作用是扩展图像的边界；d描述了卷积核在图像上进行操作时，每次前进的距离，即步长。卷积核的步长配置决策对输出特征图的尺寸有着显著的导向作用，通过调节步长，不仅可以自主操控特征图的最终规格大小，还可以间接调控卷积计算的总体负载，例如，增大步长，特征图的尺寸会相应减小，从而有效减轻计算负担；反之，若减小步长，特征图尺寸会相应增加，但这也意味着计算量会相应提升。通常情况下，步长常取值为 1 或 2。而当卷积核的大小与输入图像的尺寸无法匹配时，往往采用增加图像边框填充的策略来调和匹配。假如输入图像大小为 4×4，卷积核大小为 2×2，采用步长为 1 且不填充，那么在卷积核遍历整个输入图像后，仅能得到一个 3×3 的缩小版的特征图。在这种模式下，通过卷积操作处理后的图像逐渐变小，其中内部像素会计算多次，而边缘像素只参与一次计算，这导致图像边角的信息受到忽略与损失。因此，为了防止这类信息损失，保持图像完整性，通常采取对原始图像执行周边填充的技术手段。

2.2.2 池化层

在卷积神经网络（CNN）的架构中，池化层有着举足轻重的作用，主要用于缩减特征图的空间维度，同时确保关键特征信息的完整保留[46-47]。在实施过程中，池化层将输入的特征图细致地划分为多个互不重叠的小块区域。接着，针对每一个区域，池化层会进行一番统计分析的魔法。常见的操作包括选取每个区域中的最大值（即最大池化，Max Pooling），或是计算区域内所有值的平均数值（即平均池化，Average Pooling）。池化之后的输出为一张重构的特征图，它是原始特征图的一个简化版，更加精炼而高效。池化层显著降低了数据的复杂程度以及神经网络内参数的数量，从而减小了计算成本，并有效减轻了过拟合现象。此外，池化层提取的特征对图像的旋转、平移或缩放等几何变换表现出了良好的不变性，这一特点极大地增强了模型对外在环境变化的适应能力和鲁棒性。最大池化以及平均池化的具体操作流程如图 2-3 所示。

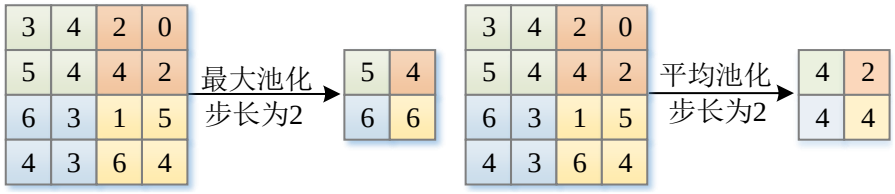

图 2-3 两种池化的操作过程

最大池化与平均池化作为两种常用的池化策略，利用减少特征图维度的方法达到降低模型参数和对抗过拟合的双重效果。在多种实践应用中，最大池化通常展现了较为卓越的效能，此技术宛如执行了一场特征的精选拔，仅从划分的小区域内挑出最显著的信号代表，以此增加模型捕捉到的非线性特性的丰度，并且雅量了深度认知的潜能。另一方面，平均池化则采取一种温和的整合方式，对特定局部区域的特征进行均匀融合，旨在平衡地保留更多的环境信息。这一策略在模型训练的后期尤为普遍，充当着“特征均质化”的角色，帮助将丰富的特征映射到一个更为紧凑、更易处理的形态上。

2.2.3 激活函数

在卷积神经网络（CNN）复杂的架构中，激活函数发挥着举足轻重的作用，它负责将卷积层精心提取的特征引入非线性变换的过程。鉴于卷积操作本身的线性本质——输入数据与卷积核的线性组合，为了让网络能识别更为复杂的模

式和特征，网络结构不得不借助激活函数来加入所需的非线性映射功能。由此，激活层利用经过挑选的激活函数，为经过卷积处理的数据赋予新的活性，使得特征图中的信息演化为表现非线性特质的形式。在众多激活函数中，常用的激活函数如 Sigmoid[48]和 ReLU[49]等。

（1）Sigmoid 激活函数：sigmoid 函数形状似 S 型曲线，是一种在两端进入饱和状态的函数。当处理各种定义域内的输入时，Sigmoid 函数能够将输入转化为介于(0,1)这一有界区间内的输出结果，赋予网络输出映射了渐变特性，Sigmoid 函数公式如式（2-5）所示。

$$f(x)=\frac{1}{1+e^{-x}} \tag{2-5}$$

其导数为：

$$f'(x)=f(x)(1-f(x)) \tag{2-6}$$

sigmoid 函数及其微分图像如图 2-4 所示。可以看出在输入为零时，sigmoid 函数梯度达到了其最大峰值，约为 0.25，输出值变化较为灵敏。当输入开始在任何方向偏离这个平衡点时，梯度逐步降低，渐渐向零逼近，会出现梯度消失的问题，这给模型训练带来阻碍。

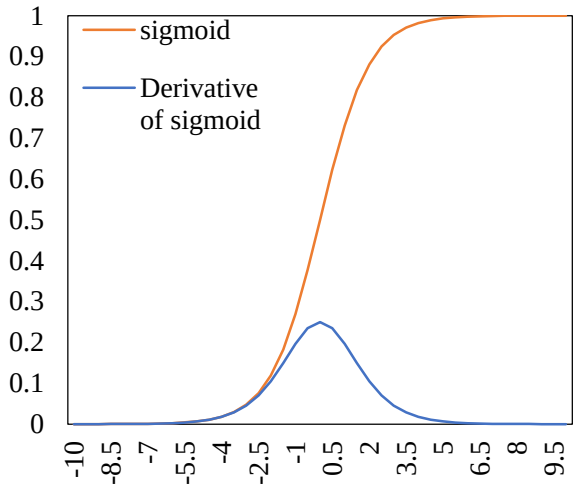

图 2-4 Sigmoid 函数及其导数图像

（2）ReLU 激活函数：ReLU 函数又被认知为修正线性单元，是一种常用的激活函数，因其结构的简约性与高效的计算特性，在众多场合下使神经网络达成快速收敛成为可能。ReLU 不仅优化了训练过程中的计算速度，而且在很大程度上缓解了深度学习模型训练中梯度消失这一通病，其计算过程如公式（2-7）所示。

$$f(x)=\begin{cases}0 & x\le 0\\ x & x>0\end{cases} \tag{2-7}$$

其导数为：

$$f'(x)=\begin{cases}0 & x\le 0\\ 1 & x>0\end{cases} \tag{2-8}$$

ReLU 函数图像及微分图像如同图 2-5 所示。可以看出，在输入值跨越零点进入正数领域时，梯度值恒为 1，能够避免梯度消失的现象，从而维持信息流的传递。当输入值位于负数范围时，梯度值将保持恒定为零的状态，这一特性使得在模型训练的过程中，部分神经元可能会陷入未被激活的状态。这些未能激活的神经元在模型的表达过程中将失去作用，无法对模型的输出结果产生影响。

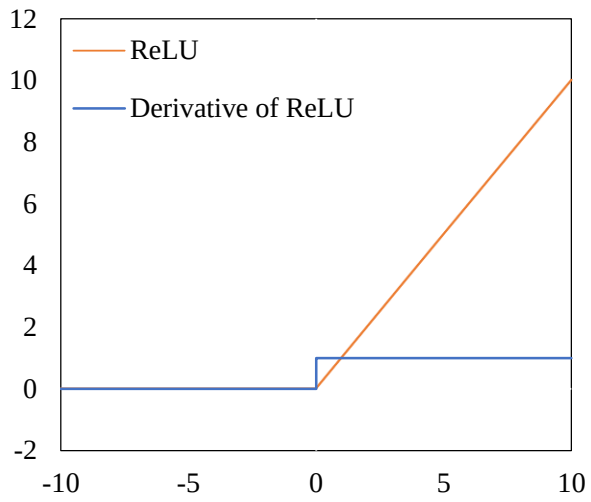

图 2-5 ReLU 激活函数及其导数图像

2.2.4 全连接层

全连接层在特征提取阶段之后负责将上层中每一神经元的信号与当前层的各个神经元连接起来，通过根据最终输出层的特定需求对高阶特征进行重塑与映射。FC 层在卷积神经网络（CNN）的架构中负责最后的特征汇总，是一种将多维的特征信息转换为二维输出的网络结构。其中，高维部分批量处理样本的细节，较低维度对应着各种任务的具体目标，这一点使得全连接层在深度学习模型中显得尤为关键。

2.3 经典网络结构

2.3.1 CenterNet 模型

CenterNet 目标检测算法提出了以点代替框进行目标检测的新思想。具体而言，该算法分为两个核心部分：目标中心定位与尺寸回归。在目标中心定位部分，通过应用高斯核生成高置信度的热力图。在尺寸回归环节，模型将每个物体的中心像素作为独立的训练样本点，从而能够直接、精确地预测出该点所对应物体的宽度和高度。此外，模型还通过预测中心偏移量的方式，精细地校正了输出步长所带来的离散误差，以实现更精准的目标定位。其模型框架如图 2-6 所示。

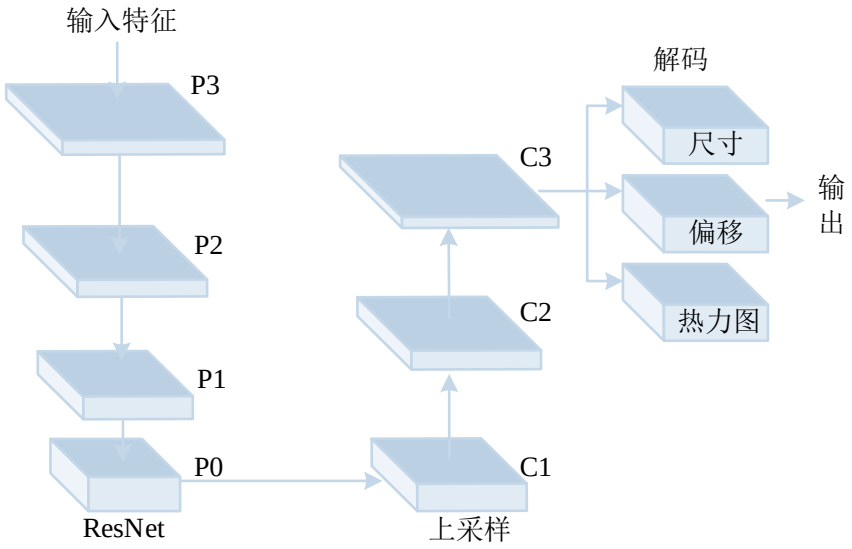

图 2-6 CenterNet 模型

本章节的CenterNet模型采用ResNet[50]作为主干特征提取网络来完成视觉特征提取的任务。ResNet 这一结构是为了解决卷积神经网络的深度加深引起的性能下降的问题而被提出，其核心在于设计了跨越数层的直连通道，通过将前层的原始输入直联至当前层的输出端，构建了残差架构。优势之处在于，网络引入残差连接的设计，使得网络在增加深度的同时，保持各层的输出在之前层输出的基础上进行有益的叠加，有效地防止信息在传递过程中的丢失。此外，在反向传播过程中，由于每一层的输出与前一层直接相关，有助于缓解梯度消失和梯度爆炸问题，这种设计策略有效提升了网络的稳定性和训练效果。假设将输入变量 x 传递至网络的某一层操作，那么该层处理后的输出可记为 $h(x)$，则

ResNet 网络的具体运算过程如式（2-9）所示。

$$f(x) = h(x) - x \tag{2-9}$$

其中，$f(x)$代表当前网络层输出与输入之间的差异，即残差。当某一层网络的输出 $h(x)$与该层接收的输入 x 相匹配时，$f(x)$为零，表示网络在此处没有额外学习到新特征。换言之，当网络输出 $h(x)$比输入 x 更接近理想输出时，$f(x)$ 的值较小。通过这种机制，每一层的网络都能通过残差学习新特征，从而有效地提升整体网络的准确性。

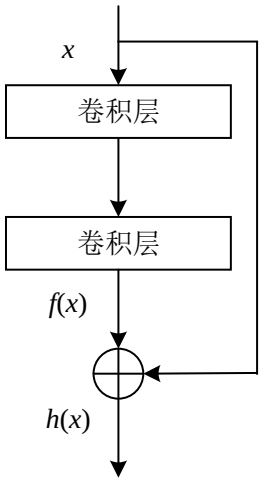

图 2-7 残差网络结构

ResNet 的网络结构由多个残差结构组合而成，每个残差结构内部不仅嵌入了多个卷积层，还引入了跨层连接结构。这一连接结构确保了特征信息在层间得以直接、高效地传递。在模型的训练过程中，这些残差单元的输出会与其自身输入融合，融合后的信息输入到后续的网络层去执行进一步的处理。ResNet 由常见堆叠层数 50、101 和 152 被分为 ResNet-50、ResNet-101 和 ResNet-152。详细网络结构参数如表 2-1 所示。

表 2-1 ResNet 网络结构

Layer name	**output size**	**50-layer**	**101-layer**	**152-layer**
Conv1	112×112	7×7，64，stride2		
Conv2_x	56×56	3×3 max pool, stride 2		
		$\begin{bmatrix}1\times1, & 64\\ 3\times3, & 64\\ 1\times1, & 256\end{bmatrix}\times3$	$\begin{bmatrix}1\times1, & 64\\ 3\times3, & 64\\ 1\times1, & 256\end{bmatrix}\times3$	$\begin{bmatrix}1\times1, & 64\\ 3\times3, & 64\\ 1\times1, & 256\end{bmatrix}\times3$

Conv3_x	28×28	$\begin{bmatrix}1\times1, & 128\\ 3\times3, & 128\\ 1\times1, & 512\end{bmatrix}\times4$	$\begin{bmatrix}1\times1, & 128\\ 3\times3, & 128\\ 1\times1, & 512\end{bmatrix}\times4$	$\begin{bmatrix}1\times1, & 128\\ 3\times3, & 128\\ 1\times1, & 512\end{bmatrix}\times8$
Conv4_x	14×14	$\begin{bmatrix}1\times1, & 256\\ 3\times3, & 256\\ 1\times1, & 1024\end{bmatrix}\times6$	$\begin{bmatrix}1\times1, & 256\\ 3\times3, & 256\\ 1\times1, & 1024\end{bmatrix}\times23$	$\begin{bmatrix}1\times1, & 256\\ 3\times3, & 256\\ 1\times1, & 1024\end{bmatrix}\times36$
Conv5_x	7×7	$\begin{bmatrix}1\times1, & 512\\ 3\times3, & 512\\ 1\times1, & 2048\end{bmatrix}\times3$	$\begin{bmatrix}1\times1, & 512\\ 3\times3, & 512\\ 1\times1, & 2048\end{bmatrix}\times3$	$\begin{bmatrix}1\times1, & 512\\ 3\times3, & 512\\ 1\times1, & 2048\end{bmatrix}\times3$
1×1	average pool, 1000-d fc, softmax			
FLOPs		3.8×10^9	7.6×10^9	11.3×10^9

其中，以 ResNet-50 为例，其框架包括五个不同的阶段，分别是 Conv1、Conv2_x、Conv3_x、Conv4_x 和 Conv5_x。初期阶段 Conv1 针对输入的图像数据进行预处理工作，运用了 7×7 的卷积层来进行全局特征的提取，并结合 BN 层的归一化处理以及 ReLU 激活函数的引入，为模型注入了必要的非线性因素。其后，利用 3×3 尺寸的最大池化层对特征图进行了下采样操作，将特征图的维度降低至 64×56×56。在 Conv2_x 至 Conv5_x 的阶段中，该网络包含了总计 16 个残差单元，每个残差单元执行了 3 次卷积操作，目的在于从先前的阶段中提取并加深特征表达。具体而言，Conv2_x 阶段中有 3 个残差单元，而 Conv3_x、Conv4_x 和 Conv5_x 阶段分别拥有 4 个、6 个以及 3 个残差单元，用于从前一阶段提取和深化特征。最后，通过采用平均池化操作和一个 1000 维度的全连接层来实现图像的分类任务。这样设计的网络结构在处理图像分类任务时表现出良好的性能。CenterNet 模型的整体结构的核心设计理念可分为以下四个部分。

（1）首先，将一张尺寸为 1×3×512×512 的图像输入到网络中通过 ResNet 主干网络的多次下采样操作来提取图像的语义信息，生成一个尺寸为 1×2048×16×16 的特征映射图。

（2）接着，利用三层反卷积模块对特征映射图进行上采样，以恢复更多的图像细节信息，得到一个尺寸为 1×64×128×128 的高分辨率特征图。随后，经过各个独立的检测头，生成描述检测目标的热力图、中心点偏移和目标尺寸信息等。

（3）再对生成的热力图进行归一化处理，通过 sigmoid 函数将其映射到 0

到 1 之间，并利用最大池化操作提取出热力图中的极大值点，同时将非极大值点置为零。然后，根据预设的置信阈值，我们筛选出前 100 个具有最高置信度的候选样本，并计算出目标的中心点位置、宽度和高度信息，以及中心点的偏移量。

（4）在得到包含中心点偏移和目标尺寸信息的候选样本后，采用解码方式结合这些信息，构出目标的检测框，最终输出整体的目标检测结果。

2.3.2 YOLOv5 模型

YOLOv5 是由 ultralytics 公司于 2020 年推出的，是一个先进的单阶段目标检测技术和开放源码的算法项目。依据算法模型的规模，YOLOv5 被分成了四种型号：小型的 s、中端的 m、大型的 l 以及超大型的 x，其在 coco 数据集[54]上的测试结论表明，随着模型规模的增大，检测性能同步提升。其网络结构由输入端、主干网络（Backbone）、特征融合网络（Neck）和输出端（head）四部分组成，如图 2-8 所示。下文将对 YOLOv5 中的重要组成部分展开详细地介绍。

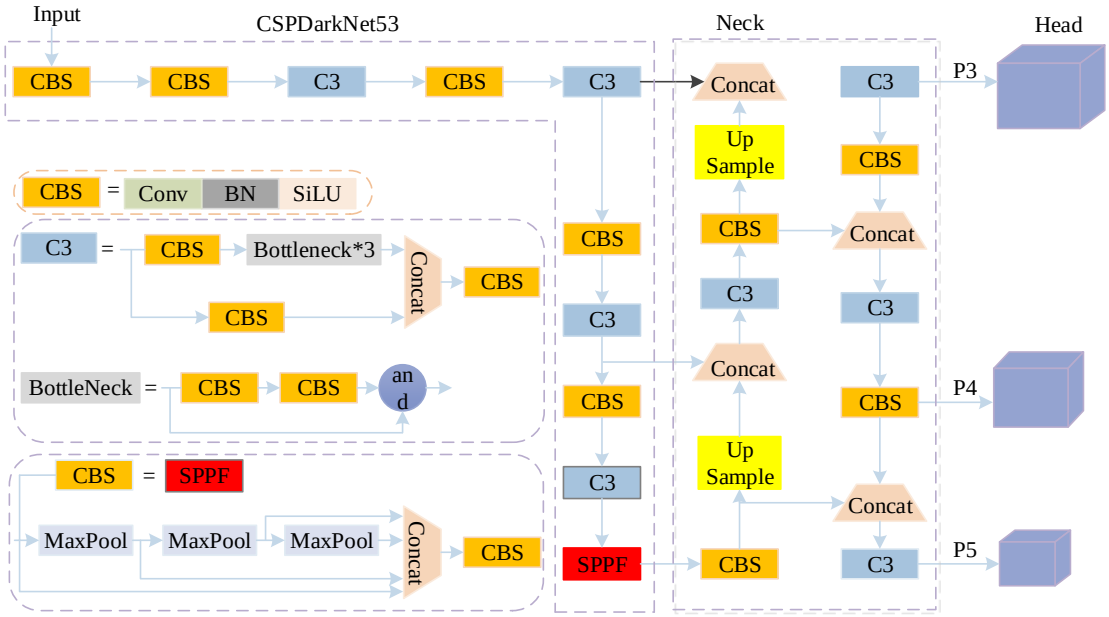

图 2-8 YOLOv5 网络框架

（1）输入端

YOLOv5 作为一种强大的实时目标检测系统，在对输入图像进行处理的过程中，整合了一系列创新的技术手段，以增强模型对于各类目标的识别与检测精度。在预处理阶段，采纳了马赛克数据扩增手法，通过将多个图像进行随机

的缩放、剪切以及拼接处理，形成新的训练样本。经过实验验证，该方法对于小目标检测的性能提升具有显著效果。随后，在锚框尺寸的确定上，引入了一种基于目标检测任务具体数据集标注情况的自适应锚框尺寸计算方法。通过这种方式，能够在模型训练初期确立最适宜的锚框大小，与标注框的差异作为反向误差传播的基础，持续对模型参数进行迭代优化。此外，还引入了自适应图像缩放技术，通过专门的计算方法减少了对图像的过度填充，这不仅在训练过程中有助于萃取更多有效特征，同时在推理过程中可以大幅减少计算量，有效提升了目标检测任务的处理速度。

（2）主干特征提取网络

YOLOv5 的架构设计中，主干特征提取网络采用了 Wang 等人[51]提出的 CSPNet 结构，并与经典的 ResNet 残差结构结合，形成了一种高效的特征提取网络。特征提取网络内部的 CSP 结构如图 2-9 所示，假设输入通道数为 c1，网络经过两次 1×1 的卷积处理，生成了具有 c_通道数的中间特征图。随后，该特征图被划分为两个分支，一个继续按照残差结构传递，另一个保持不变。最终，两个分支的特征图在融合后再次经过 1×1 卷积生成 c2 通道的输出特征图。这种结构与残差网络的残差块相结合形成了特征提取网络的主要部分 C3，即包含三次卷积操作的 CSP 结构。这种创新架构显著提升了卷积神经网络的学习效率，同时在降低计算成本和内存消耗的同时，确保了模型的精准度。

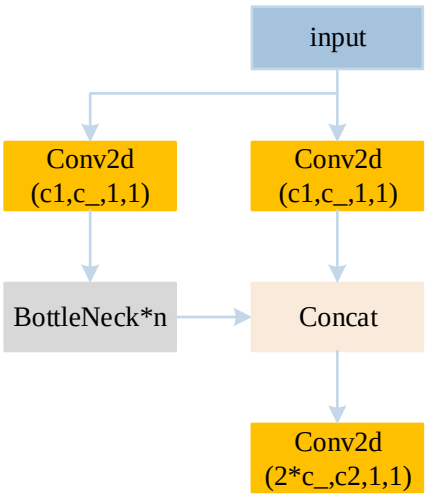

图 2-9 CSP 网络结构

（3）特征融合网络

在 YOLOv5 的深度学习架构中，特征融合网络承担着串联主干网络与输出层的关键角色，通常被称为 neck。该网络的核心作用是对各个层级的特征图进

行有效整合，以增强模型对于各种尺寸目标的识别精度。借鉴 YOLOv4 的先进设计，YOLOv5 也采用了特征金字塔网络[52](Feature Pyramid Network, FPN)代替像素聚合网络[53](Path Aggregation Network, PAN)结构的融合来完成特征图的高效融合。在其主干网络部分，多级卷积层的堆叠对输入数据执行了连贯的下采样处理，从而生成了一系列具有层次性的特征图。其中，深层特征图承载着重要的语义信息，而浅层特征图则包含了精确的定位以及清晰的轮廓信息。FPN 结构对深层特征图进行上采样且与前一级浅层特征图通过 Concat 操作进行通道方面的融合。结合 C3 模块和其他卷积操作后，FPN 在传递深层的高层次语义信息到浅层特征图的过程中，显著增强了模型对小尺寸目标的检测性能。另外，FPN 自上而下的信息流动策略旨在补偿特征提取过程中可能损失的定位细节信息，进而，YOLOv5 引进了自下而上的 PAN 结构，通过将合并后的特征图上采样并与深层的 FPN 特征图融合，以便深层特征图能够获取到更多的浅层细节信息，综合强化模型在不同尺度上的目标预测效果。与 YOLOv4 的差异在于，YOLOv5 创新性地采用了 C3 模块来替代过去的卷积结构，此举在特征融合方面的应用，为网络的特征整合力带来了显著提升，并优化了整体模型处理流程的效能。

（4）输出端

在 YOLOv5 目标检测算法的构架当中，输出层扮演着不可或缺的作用，其根本任务在于对 Neck 阶段得到的多尺度特征图进行整合性分析，并且负责完成对物体的精准定位与类别鉴别。该过程以将输入图像逐一细致划分为数个 S×S 的格点矩阵，并在每一单元格内生成诸多预选框作为潜在的检测候选。随后，算法实行非最大抑制（NMS）操作，目的是为了消除那些信任度不足的候选框，从而保障检测成果的高度精确度与稳定性。在模型训练阶段中，建立了一种检核机制，即使网络预测输出的结果与实际标注的目标框及对应的类别标注进行匹配验证，进而通过损失函数来量化两者之间的预测误差值。基于这一误差量，采取误差反向传播算法对模型中的权重参数进行调校。

2.4 评价指标

在目标检测中，评价指标用于衡量目标检测算法性能的优劣，常见的指标

包括精确率（Precision, P）、召回率（Recall, R）以及平均精确率均值（mean Average Precision，mAP）等。

精确率（P）：被分类器识别为正样本并且实际是正样本在被分类器识别为正样本中的占比，运算公式如式（2-10）所示。

$$P = \frac{T_P}{T_P + F_P} \tag{2-10}$$

其中，T_P 为正样本且预测结果是正类的数量，指的是被正确分类的正类。F_p 为负样本但预测结果是正样本的数量，代表的是被错误分类的负类。

召回率（R）：被分类器识别为正样本并且实际是正样本在所有实际是正样本中的占比，运算公式如式（2-11）所示。

$$R = \frac{T_P}{T_P + F_N} \tag{2-11}$$

其中，T_P 与式（2-10）中的含义相同，F_N 为正样本但检测结果是负样本的数量，代表的是被错误分类的正类。平均精度(average precision, AP)为 PR 曲线面积。

目标检测模型通常会用精度(mean average precision, mAP)作为评价指标。mAP 是将所有种类的 AP 取均值后的结果。

2.5 本章小结

本章对深度学习技术所涉及的基础理论进行了简要阐述，首先分析说明了深度学习卷积神经网络所涉及的基本概念；其次，介绍了两种经典网络 CenterNet、YOLOv5 的网络结构和原理；最后，讲解了目标检测算法常用的评估指标。

第 3 章 基于上下文增强和特征融合的目标检测算法

3.1 引言

目标检测作为计算视觉领域的一个基本任务，在面部分析、人数统计、行人检测、目标跟踪等领域得到广泛应用。该技术的应用对提高自动化水平、改善各行各业的效率、增强安全性和提高用户体验等方面具有重要作用。在实际的跨视域监控环境中，面对道路交通等具有多类别目标的应用场景，图像中的目标尺度、外观变化给目标检测模型提取特征信息带来困难，增加了目标识别和定位的难度。因此，尺度变化问题成为目标检测研究的重要挑战。

为解决以上问题，常见的目标检测算法研究设法利用无锚框的机制来帮助网络实现目标的分类与定位。如CenterNet网络通过将目标检测问题转化为关键点检测问题，通过热力图来预测目标的中心点，再结合宽高和偏移量来直接定位和识别目标，采用基于无锚框网络的预测方法不受预定义锚框尺寸和比例的限制，一定程度上缓解了多尺度检测的难题。但CenterNet网络仅利用单一尺度的网络架构来处理图像的特征，缺乏有效的感受野以及存在浅层位置信息丢失的问题，以致网络表征能力不足，从而影响模型的识别准确度。

针对以上问题，本章从充分利用感兴趣目标的上下文特征和减少信息损失的角度出发，提出一种基于CenterNet的目标检测改进算法。首先，在骨干网络(ResNet-50)后添加自适应上下文提取模块(Adaptive context extraction module, ACEM)，通过使用多路径空洞卷积对低分辨率特征进行不同大小感受野的特征提取，以督促网络学习多尺度上下文信息，并使用ACON-C[54]激活函数，自适应地对网络神经元进行非线性和线性之间的参数切换，加强模型的数据拟合能力。其次，提出一种注意力特征融合模块(Adaptive feature fusion module, AFFM)，通过融合高层语义特征和低层位置特征来提取不同级别特征的关键信息，从而减少信息损失，提升模型的多尺度目标识别能力。

3.2 算法流程

3.2.1 网络结构设计

在本章节，围绕多尺度目标检测任务，基于 ResNet50 作为基础网络构架，提出了一种改良型网络结构。通过设计自适应上下文增强模块（ACEM）、注意力特征融合模块（AFFM）并结合 ACON 激活函数，对原始 CenterNet 模型进行了创新性的改进，以期在目标检测的精度上取得更佳表现。所改进的网络框架如图 3-1 中所示，主要由四个部件组合而成：其一为主干特征提取网络 ResNet-50，负责对输入图像数据进行深层特征学习，并输出不同层次的特征图 P3、P2、P1 与 P0；其二为自适应上下文增强模块（ACEM），该模块对来自高层次的语义特征图 P0 提取全局上下文信息，以学习更加丰富的深层次语义；其三，注意力特征融合模块（AFFM）运用自适应加权机制，将多层特征信息进行融合，实现深层语义与浅层空间信息的高效整合；最后，该融合特征图传入到检测头（Head）输出检测结果。

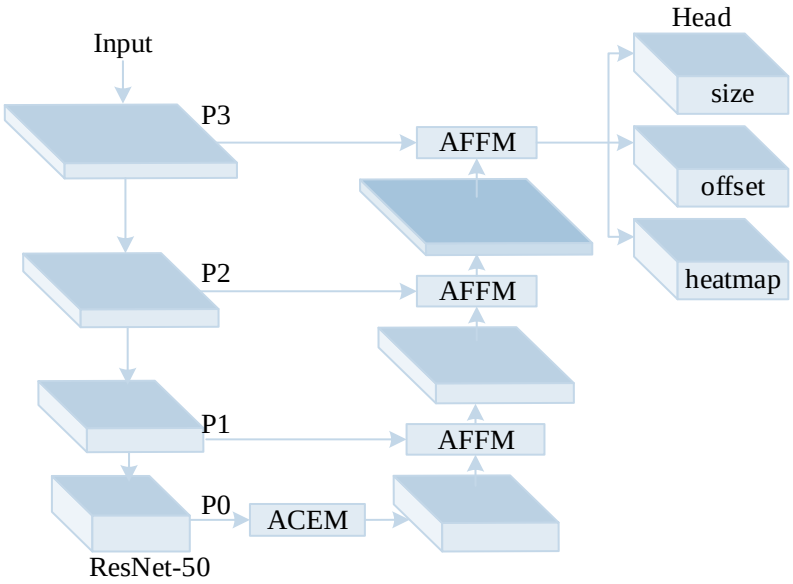

图 3-1 整体网络结构图

3.2.2 ACON 激活函数的原理

激活函数在神经网络中的作用是引入非线性特性，从而增强神经网络的学习能力。长时间以来，ReLU 激活函数一直被认为是最优秀的神经网络激活函数之一，主要因为其具有非饱和性、稀疏性等优秀特性。然而，它可能会导致神经元坏死等严重后果。本章使用 ACON 激活函数，明确地学习函数的线性与非线性，给算法提供非线性建模能力。其原理如式（3-1）所示。

$$f_{\text{ACON}}(\eta_{\text{a}}(x),\eta_{\text{b}}(x))=(\eta_{\text{a}}(x)-\eta_{\text{b}}(x))\times\sigma[\beta\times(\eta_{\text{a}}(x)-\eta_{\text{b}}(x))]+\eta_{\text{b}}(x) \quad (3\text{-}1)$$

其中，$\eta_{\text{a}}(x)$和$\eta_{\text{b}}(x)$为输入样本，$\sigma(\bullet)$为 sigmoid 函数，β为平滑因子，通过改变平滑因子β的大小可调节其渐进上界和下界的速度。$\eta_{\text{a}}(x)$和$\eta_{\text{b}}(x)$在取不同的值时对应不同的激活函数，当$\eta_{\text{a}}(x)=x$，$\eta_{\text{b}}(x)=0$时，对应 ACON-A 激活函数；当$\eta_{\text{a}}(x)=x$，$\eta_{\text{b}}(x)=px$时，对应 ACON-B 激活函数；当$\eta_{\text{a}}(x)=p_1x$，$\eta_{\text{b}}(x)=p_2x$时，对应 ACON-C 激活函数。详细运算过程如式（3-2）、式（3-3）和式（3-4）所示。

$$f_{\text{ACON-A}}(x)=x\times\sigma(\beta\times x) \quad (3\text{-}2)$$

$$f_{\text{ACON-B}}(x)=(1-p)x\times\sigma[\beta\times(1-p)x]+px \quad (3\text{-}3)$$

$$f_{\text{ACON-C}}(x)=(p_1-p_2)\times\sigma[\beta\times(p_1-p_2)x]+p_2x \quad (3\text{-}4)$$

其中，p、p_1、p_2为待训练参数，用于调节学习的上下界。ACON-C 激活函数通过改变p_1和p_2两个参数的大小可分别调节学习的上界和下界，并配合平滑因子β来调节其渐进上界和下界的速度，具有更好地数据拟合能力，效果优于 ACON-A 和 ACON-B 激活函数。因此，本章采用 ACON-C 激活函数，并应用在自适应上下文提取模块（ACEM）中。

3.2.3 自适应上下文提取模块

如果模型拥有较广的感受野，它将能够获取更为丰富的全局特征信息。这些全局上下文特征对于提升模型检测图像中多尺度的性能具有关键性作用。在空洞卷积技术问世之前，研究人员常通过下采样的方式来扩展感受野，但这样做往往会导致大量重要信息流失，并显著降低了特征图的分辨率。2017 年，Fisher Yu 等人[55]在图像分割的研究领域内引入了具有划时代意义的空洞卷积手段，有效解决了该问题。这种扩张卷积是在传统卷积中加入了填充操作，通过增大卷积核的尺寸的方式，实现了扩展特征图分辨率的目标。这种操作优化了全局信息的整合效果，扩大了感受野范围，成功应对了不同尺度目标对感受野的需求。但研究指出，单一尺度的空洞卷积可能产生所谓的格化效应[56]，当空洞率过高时，卷积核内部空白像素增多，造成了卷积操作的稀疏化，这样就难

以捕捉到图像中的关键信息，影响检测效果。为了解决这一问题，本章设计了一种创新的自适应上下文提取模块（ACEM），通过整合多个不同扩张率的卷积层产生的特征图，有效提升了网络获取全局特征的能力，并有效补偿了由空洞卷积导致的格化效应。结构如图 3-2 所示，X 代表特征输入。

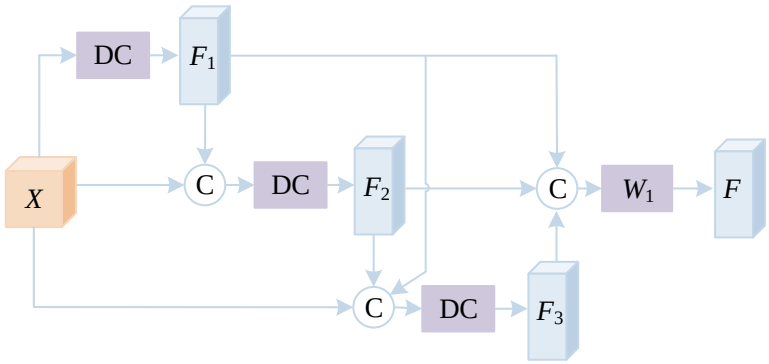

图 3-2 自适应上下文提取模块 ACEM

首先，利用三层并行的空洞卷积(dilated convolution, DC)模块对特征图进行多尺度特征采集，并在 DC 模块中采用 ACON-C 激活函数，自适应地控制网络的线性和非线性程度，以此获取有效的上下文信息用于遮挡目标检测，DC 模块是由 1×1 卷积和不同速率(rate=3,6,12)的空洞卷积组合而成，并且三层并行的 DC 模块采用密集连接方式（即每一层输出特征 F_i 和输入 X 拼接后再传入下一层）进一步扩大感受野，来获得更丰富的多尺度上下文信息。多尺度特征 F_1、F_2、F_3 提取的过程如式（3-5）、式（3-6）、式（3-7）所示。

$$F_1 = D_{r=3}(X) \tag{3-5}$$

$$F_2 = D_{r=6}(\text{cat}(X, F_1)) \tag{3-6}$$

$$F_3 = D_{r=12}(\text{cat}(X, F_1, F_2)) \tag{3-7}$$

其中，$D_{r=i}(\bullet)$ 代表 DC 模块，r 代表的是扩张率，$\text{cat}(\bullet)$ 表示沿通道维度拼接。接着，将多尺度特征 F_1、F_2、F_3 沿通道维度拼接得到多尺度特征图 $\text{cat}(F_1, F_2, F_3)$，最后，通过 1×1 卷积模块融合粗粒度和细粒度特征，详细运算过程如式（3-8）所示。

$$Z = W_1(\text{cat}(F_1, F_2, F_3)) \tag{3-8}$$

其中，$\text{cat}(\bullet)$ 表示沿通道维度拼接，$W_1(\bullet)$ 代表 1×1 卷积运算。

3.2.4 注意力特征融合模块

经 CenterNet 主干特征提取网络（ResNet-50）处理后可得到 4 种尺度的特征。低层次特征含有丰富的空间定位和细节纹理信息，然而其携带的语义信息较为贫乏。与此相反，高层次的特征则包含了更多语义信息，对于物体的描述有着更强的表达能力，但对目标细微特征的捕捉能力有限。若仅依赖于深层次的高层特征，就会忽略低层次的重要信息。基于此，为了综合不同深度的特征信息，需要执行将深层和浅层信息融合的过程，但该方法存在不同层次信息间相互干扰的问题。为此，提出了一种嵌入注意力机制的特征融合模块（AFFM），该模块能够对各层信息进行加权组合。通过这种策略，可以有效地融合并发挥有益信息的作用。借助该模块的实施，在确保低层详尽的位置与细节特征的同时，也加强了模型对目标高层语义信息的学习力度，从而提升模型对不同尺度特征的提取能力。

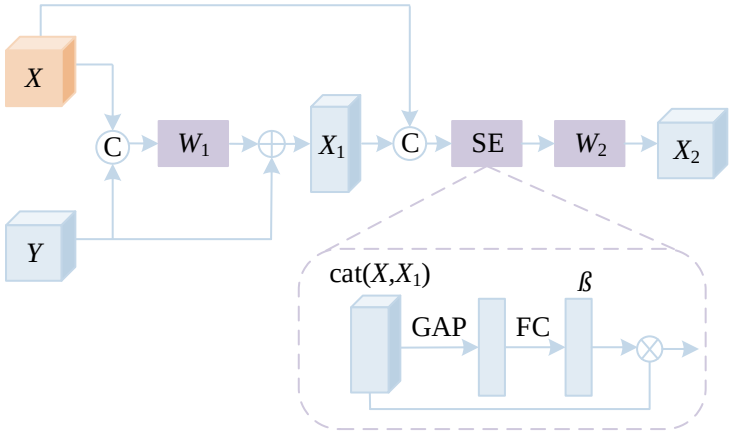

图 3-3 注意力特征融合模块 AFFM

结构如图 3-3 所示，高阶语义特征图和低阶语义特征图作为输入，标记为 X 与 Y。首先，通过在通道方向上将特征 X 和 Y 连接得到特征图 $\mathrm{cat}(X,Y)$。其次，特征图 $\mathrm{cat}(X,Y)$ 传入 1×1 卷积和 3×3 卷积实施通道的压缩并整合跨通道信息获得特征图 $W_1(\mathrm{cat}(X,Y))$，为了避免梯度消失干扰检测效果，将结果与原始特征融合，输出一个初步的合成特征图 X_1。具体计算公式如式（3-9）和式（3-10）所示。

$$W_1(\mathrm{cat}(X,Y)) = W_{3\times3}(W_2(\mathrm{cat}(X,Y))) \tag{3-9}$$

$$X_1 = Y \oplus W_1(\mathrm{cat}(X,Y)) \tag{3-10}$$

其中，$W_2(\bullet)$ 代表 1×1 卷积运算，$W_{3\times3}(\bullet)$ 代表 3×3 卷积运算，$\mathrm{cat}(\bullet)$ 表示在通道维度上拼接，$\oplus$ 表示按元素加合。

再次，沿通道层次上将特征 X_1 和 X 连接得到特征图 $\text{cat}(X_1, X)$。然后，特征图 $\text{cat}(X_1, X)$ 在压缩激励网络[57](squeeze-and-excitation networks, SE)中自动适应地融合有用的信息，SE 是通过全局平均池化技术来获得通道层次的信息，并利用全连接层(fully connected, FC)实现通道的数据交互，从而得到融合权重 β，再将融合权重 β 和特征图 $\text{cat}(X_1, X)$ 按元素相乘，输出加权融合的特性图 $\text{cat}(X_1, X) \odot \beta$。最后，特征图 $\text{cat}(X_1, X) \odot \beta$ 输至 1×1 卷积，实现通道的压缩，从而输出特征图 X_2。具体计算公式如式（3-11）和式（3-12）所示。

$$\beta = \text{FC}(\text{GAP}(\text{cat}(X_1, X))) \tag{3-11}$$

$$X_2 = W_2(\text{cat}(X_1, X) \odot \beta) \tag{3-12}$$

其中，$W_2(\bullet)$ 代表 1×1 卷积运算，⊙表示按元素相乘，$\text{cat}(\bullet)$ 表示沿通道维度拼接。

3.3 实验设置与结果分析

3.3.1 实验设置与数据集

本章实验环境配置表 3-1 所示，操作系统为 Windows10，CPU 为 12 核 Intel(R) Xeon(R) Platinum 8255C CPU @ 2.50GHz，内存 43G，显卡 Nvidia GeForce RTX 2080Ti(11GB)，CUDA 版本为 10.0，深度学习框架为 pytorch1.2.0。

表 3-1 实验环境配置

实验配置	型号或参数
系统	Windows10
CPU	12 vCPU Intel(R) Xeon(R) Platinum 8255C CPU @ 2.50GHz
GPU	Nvidia GeForce RTX 2080Ti（11G）
CUDA	10.0
深度学习框架	Pytorch1.2.0
算法语言	Python

在本章研究中，所采纳的算法以 512×512 像素的分辨率输入进行模型训练。训练过程设定为 100 个周期，采取固定比例递减策略对学习率进行调控。同时，借助 CenterNet 模型预训练的参数进行权重初始化，进而加速模型在执行推断任务时的处理效率及算法的收敛速度。在训练的最初 50 个周期内，模型主要对特

征抽取网络部分的参数实施冻结训练，训练批次为 8，初始学习率设定为 0.001，并采取在每个训练周期结束时将学习率减少为原来的 0.94 倍。随后，在第 51 至 100 个周期的训练阶段，解冻参数后，执行全面训练，此时批量大小调整为 4，初始学习率设置为 0.0001，维持学习率在每个结束的训练周期后以 0.94 的比例递减。实验参数设置如表 3-2 所示。

表 3-2 实验参数设置

迭代次数	批次	学习率	权重衰减率
0~50	8	0.001	0.94
50~100	4	0.0001	0.94

本章实验采用了 PASCAL VOC 数据集[58]对研究模型的识别性能进行了全面的测试。该数据集囊括了 20 个类别的各类物体，由 VOC2007 和 VOC2012 联合训练集构成，总共提供了 16551 幅图像和 40058 个标注目标作为训练样本。为了验证模型的检测能力，采用的数据集包括了 4952 张图像和 12032 个目标作为测试样本，每幅图像平均包含大约 2.4 个目标。其数据集组成具体如下表 3-3 所示。

表 3-3 PASCAL VOC 数据集训练和测试数据统计

Data	Trainval		Test	
	Images	Objects	Images	Objects
VOC2007	5011	12608	4952	12032
VOC2012	11540	27450	0	0
Total	16551	40058	4952	12032

3.3.2 消融实验

本章中，在相同的实验设定以及 VOC 数据集下进行消融实验，来评估所提方法的有效性，实验数据如表 3-4 和 3-5 所示。

其中，表 3-4 列出了三种改进方法，分别是 ACEM、ACON-C、AFFM。其中，ACEM 表示对主干特征提取网络(ResNet-50)输出特征进行上文信息提取，ACON-C 表示在 ACEM 中加入非线性因素，AFFM 表示对不同维度的特征进行信息融合，通过不同组合方式在基线（CenterNet）上引入 ACEM、ACON-C、AFFM 这三种改进，表格所列每组实验数据中的“√”代表该组实验引入了此改进方法，“×”代表未引入该改进方法。Experiment1 表示在基线（CenterNet）

上添加ACEM，由表3-4实验结果可知，mAP得到了2.15%的提升，说明ACEM有加强网络特征提取能力的作用，能有效聚合上下文信息；Experiment2 表示在Experiment1 的基础上对特征进行非线性处理，与 Experiment1 方法相比，mAP提升了 0.9%，说明 ACON-C 激活函数有提升网络数据拟合性能的作用；Experiment3表示在基线（CenterNet）上添加AFFM，与基线算法相比，mAP得到了 1.42%的提升；Experiment4 表示在 Experiment1 的基础上添加 AFFM，与Experiment1方法相比，mAP提升了0.37%，说明AFFM有丰富和细化多尺度特征的作用，能提取对识别任务有用的关键特征；Experiment5表示在Experiment2的基础上再对不同粒度的特征进行加权融合，与Experiment2方法相比，mAP增加了 0.67%，使得目标检测算法在 PASCAL VOC 数据集上的性能得到进一步提升。

表 3-4 VOC 2007 测试集上的消融实验结果(%)

Method	**ACEM**	**ACON-C**	**AFFM**	**mAP**
CenterNet	×	×	×	80.10
Experiment1	✓	×	×	82.25
Experiment2	✓	✓	×	83.15
Experiment3	×	×	✓	81.52
Experiment4	✓	×	✓	82.62
Experiment5	✓	✓	✓	**83.82**

表 3-5 不同类别的检测精度对比(%)

Category	**CenterNet**	**Experiment1**	**Experiment2**	**Experiment3**
aeroplane	88.28	89.83	90.26	92.69
bicycle	87.61	89.53	89.59	90.03
bird	79.00	81.90	82.71	84.51
boat	70.63	73.57	74.49	76.12
bottle	63.91	65.03	65.97	69.09
bus	87.00	89.69	90.78	90.74
car	90.03	91.03	91.13	91.41
cat	90.20	91.57	91.36	91.94
chair	64.18	64.50	65.92	67.58
cow	86.43	90.31	89.93	88.29
dining table	71.75	72.33	77.47	78.12
dog	85.18	87.39	88.49	89.41
horse	89.61	90.24	91.33	89.63

motorbike	87.71	89.67	89.76	90.79
person	85.15	85.49	85.83	86.77
potted plant	48.33	54.88	56.80	55.74
sheep	82.60	87.10	88.06	87.43
sofa	77.09	79.81	80.09	81.07
train	86.08	87.22	88.39	89.59
tvmonitor	81.33	83.93	84.56	85.42
mAP	80.10	82.25	83.15	83.82

从表3-5中的每个类的检测精度数据去分析可以得到以下结论：大量密集的实例在空间定位上产生相互覆盖的状况会增加检测的复杂度。当在基线网络中添加 ACEM 后，并使用 ACON-C 激活函数对特征进行非线性处理，改进后的算法在复杂环境下对物体的检测效果要优于之前的算法。例如在 bird（鸟）、cow（牛）、sheep（羊）这些易于成群出现的复杂场景中的目标 AP 分别有 3.71%、3.5%、5.46%的提升，在 dining table（餐桌）、potted plant（花盆）这些边缘模糊易出现遮挡的目标 AP 分别有了 5.72%、8.47%的提升。自适应上下文提取模块（ACEM）旨在借助被测物体周围的有效上下文信息，在不同的感受域中捕获上下文信息，挖掘标签中相互依赖的语义信息，从而提升检测性能；对于网络在进行下采样卷积操作时造成的目标位置信息的丢失，在 Experiment2 的基础上加入注意力特征融合模块（AFFM）后，模型的位置定位能力得到了改善，且改进后的算法在检测多种类别的性能上都有所提升。例如在 bird（鸟）、bottle（杯子）这些类别 AP 分别又有了 1.80%、3.12%的提升，特征融合网络旨在将不同尺度信息相融合，弥补下采样操作过程中丢失的位置信息，从而提升模型的多尺度表征能力；在基线模型中同时引入自适应上下文提取模块（ACEM）、ACON-C 激活函数和注意力特征融合模块（AFFM），可以有效提升网络对多个类别的检测性能。例如，像 dining table（餐桌）和 potted plant（花盆）等物体的 AP 分别提升了 6.37%和 7.41%，使改良后的模型的 mAP 提高至 83.82%。

3.3.3 与其他方法进行比较

为了验证本章算法的有效性，对比实验在公共数据集PASCAL VOC数据集进行，将本章所提算法与现有的目标检测先进算法进行比较，其主要包括Faster R-CNN[27]、CoupleNet[59]、SSD[35]、RefineDet[60]、RFBNet512[61]、CenterNet[43]、

YOLOv3[36]、FCOS[45]、文献[62]、文献[63]、文献[64]等，具体实验数据如表3-6所示。

表 3-6 VOC2007 测试集和其他方法对比实验结果(%)

Method	**Backbone**	**mAP**
Faster R-CNN[27]	ResNet-101	76.4
CoupleNet[59]	Reaidual-101	82.7
SSD[35]	VGG-16	74.3
RefineDet[60]	VGG-16	81.8
RFBNet512[61]	VGG-16	82.2
CenterNet[43]	ResNet-50	80.1
CenterNet[43]	ResNet-101	78.7
CenterNet[43]	DLA34	80.7
YOLOv3[36]	DarNet53	80.3
FCOS[45]	ResNet-50	78.7
文献[62]	CBResNet101	83.2
文献[63]	Res101-FcaNet	82.3
文献[64]	ResNet-50	74.7
Ours	**ResNet-50**	**83.8**

本章提出的基于上下文增强和特征融合的目标检测算法与基线(CenterNet)算法相比，在 VOC2007 测试集上 mAP 增加了 3.7%。由于 CenterNet 算法采用的网络结构表征能力不足，缺乏有效的上下文信息和目标空间位置信息，导致遮挡场景中目标识别准确率低，本章通过构建 ACEM 和 AFFM、引入 ACON 激活函数等三方面改进原始 CenterNet 网络，挖掘深层特征的上下文语义信息，并将深层特征的上下文语义信息和多维度的空间位置信息进行加权融合，加强网络对有用信息的关注度，从而提升目标检测性能。本章算法与 Faster R-CNN 和 CoupleNet 算法相比，mAP 分别增加了 7.4%和 1.1%，Faster R-CNN 和 CoupleNet 算法利用单一尺度特征生成基于候选区的局部特征，缺乏本章对目标全局上下文信息的关注；与 SSD 算法相比，mAP 增加了 9.5%，SSD 算法使用不同分辨率的特征图去预测目标，缺乏本章对目标空间位置信息的关注；与 FCOS 算法相比，mAP 增加了 5.1%，FCOS 算法利用特征金字塔对不同层次特征进行融合，没有考虑到不同层次特征间的关联性，本章对不同特征图通道间的关联性进行建模，实现不同分辨率特征的充分融合；与在同一基线（CenterNet）算法上改进的文献[63]相比，mAP 增加了 1.5%，文献[63]构建轻量级的信息融合网络减

少了网络参数，但丢失了目标的细腻特征，并且没有关注到不同层次特征间的相关性，本章算法使用注意力特征融合模块(AFFM)更能处理图像的细节，又通过学习不同层次特征间的语义关联性对多层特征加权融合，进一步加强了网络对细节信息的捕获能力。

通过实验结果表明本章算法对目标多尺度上下文特征和空间位置特征提取效果更好，更能合并对目标识别有利的信息，更能应对复杂环境下的目标检测任务，进一步证明了所提算法的有效性。

3.3.4 可视化结果

为了更直观地评价本章算法，将基线模型和改进后的模型分别在 VOC2007 测试集上测试。识别结果如图 3-4 所示。第一行图像为基线算法检测结果，第二行图像为改进算法的检测结果。

Baseline

Ours

图 3-4 可视化检测结果对比图

从图中对比结果可以看出，在原始模型中，一些目标容易出现误检和漏检。但经过改进后，这些问题得到了缓解。从第一组羊群对比图和第四组花盆对比图可以看出，在密集场景下（图中被测物体之间相互遮挡），基线模型检测性能不佳，出现了漏检的现象，而本章算法采用了自适应上下文提取模块(ACEM)对基线模型进行改进后，借助被测物体的上下文信息，有效地解决了漏检的问题，同时提高了羊的检测框的置信度得分。从第二组室内家居环境的结果对比可以看出，基线模型将电视机里的人误判为检测对象，而本章算法通过采用 ACON-C 激活函数对特征进行非线性处理，具有良好的数据拟合能力，对于正样本表

现为非线性拟合，重组后的特征加强了正样本的表征能力，而在负样本中变现为线性拟合，重组后的特征减弱了背景特征的干扰，降低了模型的误检。从第三组道路交通场景中的对车和人的识别结果对比可以看出，原算法存在漏检的问题且目标的定位缺乏精确性，本章提出的算法，通过整合注意力特征融合模块，将高级语义信息和底层位置信息混合在一起，大幅提升了车辆的定位精准度，同时准确地识别出了公交车里的人。从第四组的结果对比可以看出，原算法出现漏检和定位不准的问题，而本章不仅减少了误检，对船和人的定位也更准确。

3.4 本章小结

本章提出了一种基于上下文增强和特征融合的目标检测算法。针对 CenterNet 无锚框目标检测算法的特征提取能力不佳的问题，提出了自适应上下文提取模块和特征融合模块进行改进。前者通过挖掘深层网络的多感受野特征，学习目标周围的上下文信息，增强了网络对全局上下文特征的表达能力，并联合 ACON 激活函数，加强了网络对复杂数据的拟合性能。后者对深层网络的上下文语义信息和不同维度的空间位置信息进行融合，提高了网络对多尺度特征信息的捕获能力，最终通过在公开数据集上的实验结果表明了所提方法的有效性。本章算法与基线 CenterNet 算法相比，在 PASCAL VOC 数据集下 mAP 高出 3.71%，有效提升了网络识别准确度，可应用于道路交通、智能家居等多类别目标领域，具备精准的检测以及稳定的识别效果，后续章节将继续优化模型，尝试特征融合能力更强的网络结构。

第 4 章 基于多感受野注意力和特征融合的目标检测算法

4.1 引言

面对多类别目标的应用场景，论文第三章的算法通过提取上下文信息和融合多级特征的方式来丰富多尺度特征和减少位置信息损失，有效地提升了算法的性能，但是，受限于单一尺度的网络结构，仅考虑融合不同层次特征忽略不同尺度目标特征的完整表达对输出分辨率大小的需求，容易导致细节信息的丢失，使目标识别准确度下降。常用的解决方法是利用多个独立尺度的预测头来帮助网络实现物体的分类与定位。如 YOLOv5s 网络利用特征金字塔和特征融合的思想输出多个检测头在不同尺度的特征图上进行预测，该方法可以有效地缓解由固定尺度特征引起的重要目标信息丢失的问题，但缺乏对不同尺度特征之间关联性的学习。

因此，针对上述问题，本章提出了一种基于多感受野注意力和特征融合的目标检测改进算法。从弱化相关目标干扰和减少关键信息丢失的角度出发，对 YOLOv5s 模型进行改进，以达到有效提取特征以及最大化融合有用信息的目的。首先，在主干特征提取网络之后添加多感受野注意力模块，通过并行的不同大小卷积核来获取高层网络的不同尺度特征，督促模型学习多尺度信息，同时利用注意力结构来学习图像中不同尺度特征的关联性，以确保适当的特征权重分配，从而实现多尺度特征提取的有效性。其次，采用重新参数化的广义特征金字塔融合网络 RepGFPN 代替颈部特征融合网络对不同层次的特征信息进行合并，通过跳级连接和重参数化的方式来挖掘上下层以及相邻层的特征信息，合并后的特征再传递到不同尺度的输出层，以减少重要信息的丢失，从而实现高级语义信息和低级空间信息的充分交换，提升模型的多尺度目标识别能力。

4.2 算法流程

4.2.1 网络结构设计

本章网络框架基于 CSPDarkNet53 骨干网络提出适用于多尺度目标检测算法，

通过引入 SK 多感受野注意力模块及 RepGFPN 对原始的 YOLOv5s 模型进行了改进，以提升目标检测精度。改进后的网络结构如图4-1所示，由主干特征提取网络(CSPDarkNet53)，多感受野注意力模块（SK），重新参数化的广义特征金字塔网络(RepGFPN)和检测头(Head)四个部分组成。具体来讲，CSPDarkNet53 主干网络负责深层特征的提取过程，对输入的图像数据进行处理，并输出多层次信息的特征图 F2、F1 和 F0。基于此，SK 多感受野注意力模块对主干网络提取出的高层语义特征图 F0 的不同感受域特征进行学习，利用动态权重调整策略，灵活地提取图像中至关重要的多尺度特征，并进一步输出了加强了特征表现能力的多尺度表达图 P0。随后，采用重新参数化的广义特征金字塔网络（RepGFPN）对网络进一步优化，将 CSPDarkNet53 输出的深层特征图 F2、F1 以及 SK 模块强化后的 P0 多尺度特征进行了深入的信息合并。通过跨层连接结构和重参数化技术融合不同尺度级别下的特征信息，实现由高层次的语义信息到底层空间信息的有效整合。最后，该融合特征图传入到检测头(Head)输出检测结果。

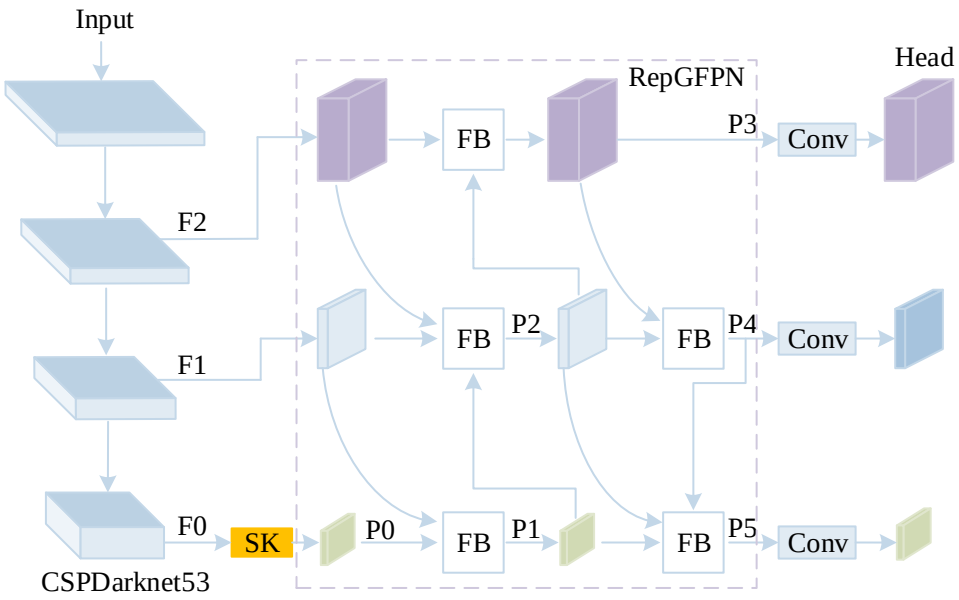

图 4-1 总体网络结构

4.2.2 SK 注意力模块

为了进一步适配卷积神经网络中神经元感受野的特性，并最大限度地挖掘大感受野所能提供的全局上下文信息，同时考虑到不同尺度目标的检测效果具有显著差异，受到 SKNet[65]网络的启发，在 YOLOv5 模型的深度特征提取阶段

设计了一种可灵活调节感受野的通道注意力机制——SK 注意力。SK 注意力模块动态地利用多种尺寸的卷积核提取不同感受野下的特征图，以学习网络通道信息。旨在通过多尺度特征提取扩展网络的感受野，捕获多层次丰富的目标信息。此外，该模块通过平衡不同尺度特征图的注意力权重，自适应地对各维度特征施加不同的关注强度，从而有效加强对图像中重要信息的辨识和捕获能力，以凸显对检测任务关键的图像细节。这样的策略不仅拓宽了特征表示的维度，亦显著提高了特征的表征力，综合这些技术的应用实现对多尺度目标的准确检测。

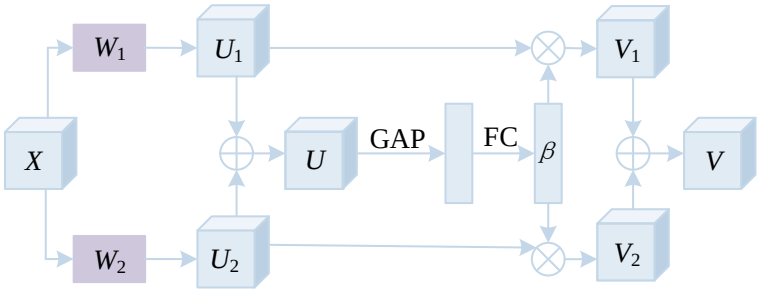

图 4-2 SK 注意力模块

由图 4-2 可知，SK 注意力模块主要分为三部分：分割、融合和选择。在分割阶段，使用一个 3×3 的卷积和一个 5×5 卷积来分离输入向量 X，通过学习输入特征 X 的不同感受野的特征信息，以输出多尺度特征 U_1 和 U_2，具体运算过程如式（4-1）、式（4-2）所示。

$$U_1 = W_1(X) \tag{4-1}$$

$$U_2 = W_2(X) \tag{4-2}$$

其中，$W_1(\bullet)$ 代表 3×3 卷积运算，$W_2(\bullet)$ 代表 5×5 卷积运算。

在融合阶段，首先，在通道方向上将特征 U_1 和 U_2 连接得到初步融合特征图 U。然后，使用全局平均池化技术(gloabe average pooling, GAP)处理特征图 U 获得通道层次的信息，从而建立信道之间的依赖性。最后，利用全连接层(fully connected, FC)来学习拟合通道之间的复杂关联，以实现通道间的数据交互，使通道之间的关系具有灵活性和非线性，最终的输出为权重值 β，计算过程如式（4-3）所示。

$$\beta = \mathrm{FC}(\mathrm{GAP}(U_1 + U_2)) \tag{4-3}$$

选择阶段是一种简单的加权操作。将融合阶段计算得到的权重值 β 乘回分

裂阶段的输出特征得到 SK 注意力模块的最终输出特征图 V，通过学习不同的信道化场景的差异性来加强学习目标的有用信息，增强网络的尺度感知力，并对加权后的特征图进行合并，得到多尺度特征图 V，具体计算过程如式（4-4）所示。

$$V = \beta * U_1 + \beta * U_2 \tag{4-4}$$

4.2.3 重新参数化的广义特征金字塔网络 RepGFPN

YOLOv5 在颈部采用 PANet 融合多尺度特征映射，通过增加额外的自底向上路径，利用低层中准确的定位信息来增强整个特征层次，以缩短高层与低层之间的信息传播路径[66]。然而，该设计存在自底向上路径缺乏高层语义信息与低层空间信息充分交互的问题，从而影响多尺度目标检测效果。为了解决这一问题，本章采用 RepGFPN[67]作为颈部结构，具备跳跃连接和跨尺度融合的高效层聚合网络，用以构建模型颈部的多尺度输出结构。RepGFPN 对特征融合的拓扑架构进行了全方位的优化，并引入了基于重参数化思想的层聚合连接结构的融合块 Fusion Block，FB），致力于实现不同尺度特征之间的有效融合，为不同尺度特征赋予合适的通道数量。RepGFPN 在不增加额外的计算负担的情况下取得了更高的精度，在融合模块 FB 的结构示意图中，如图 4-3 所示。结构方面，以 P4 层为例，在第一步，P1 经过 2 倍上采样并与 F1、F2 融合生成中间节点 P2；接着，P3 经过 2 倍下采样与 P2 融合生成 P4，同理得到 P5；最后，P2 经过 2 倍上采样与 F2 融合生成 P3。

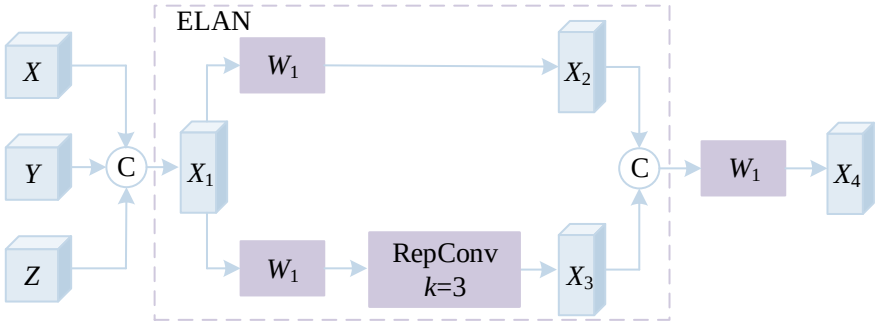

图 4-3 融合模块 FB

由图 4-3 可知，输入特征 X、Y 和 Z 通过 FB 组件实现了特征的融合。首先，在通道方向上连接特征 X、Y 和 Z 获得初步融合特征图 X_1。其次，特征图 X_1 通过集成了高效层聚合(efficient layer aggregation networks, ELAN)[68]结构和重参数

化卷积结构(RepConv)来增强特征表达能力，从而输出特征图 X_2 和 X_3，FB 的具体运算过程如式（4-5）、式（4-6）、式（4-7）所示：

$$X_1=\text{cat}(X,Y,Z) \tag{4-5}$$

$$X_2=W_1(X_1) \tag{4-6}$$

$$X_3=W_2(W_1(X_1)) \tag{4-7}$$

其中，$W_1(\bullet)$ 代表 1×1 卷积运算，$W_2(\bullet)$ 代表 3×3 的 RepConv 模块。最终，在通道方向上连接征 X_2 和 X_3，并使用 1×1 卷积来调整通道维数，以实现多尺度特征的进一步融合，从而输出融合特征图 X_4，具体运算过程如式（4-8）所示。

$$X_4=W_1(\text{cat}(X_2,X_3)) \tag{4-8}$$

4.3 实验结果

4.3.1 实验设置

本章具体实验环境配置如表 4-1 所示，其中，操作系统为 Windows10，CPU 为 12 核 Intel(R) Xeon(R) Platinum 8255C CPU @ 2.50GHz，内存为 43G，显卡为 Nvidia GeForce RTX 2080Ti(11GB)，CUDA 版本为 11.1，深度学习框架为 pytorch1.9.0。

表 4-1 实验环境配置

实验配置	型号或参数
系统	Windows10
CPU	12 vCPU Intel(R) Xeon(R) Platinum 8255C CPU @ 2.50GHz
GPU	Nvidia GeForce RTX 2080Ti（11G）
CUDA	11.1
深度学习框架	Pytorch1.9.0
算法语言	Python

设置的具体实验参数如表 4-2 所示，在本章研究中所采纳的算法以 512×512 像素的分辨率作为输入进行每轮的模型训练，训练过程设定 50 个训练周期，训练批次设定为 16，优化器采用 Adam 优化器，设定的起始学习率为 0.01，设定的权重衰减为 0.0005。实验的数据集也采用上一章节的即 PASCAL VOC 数据

集。

表 4-2 实验参数设置

迭代次数	批次	优化器	初始学习率	权重衰减
50	16	Adam	0.01	0.0005

4.3.2 消融实验

本章中，在相同的实验设定以及 VOC 数据集下进行消融实验，来评估所提方法的有效性,实验数据如表 4-3 和 4-4 所示

其中，SK 表示对主干特征提取网络（DarkNet53）输出的深层特征在多个感受域进行有用特征提取；RepGFPN 表示对不同尺度的特征进行信息融合。Experiment1 表示在基线（YOLOv5）上添加 SK，由实验结果可知，mAP 得到了 1.0%的提升，说明 SK 有加强网络多尺度特征提取能力的作用，能有效聚合不同感受野信息；Experiment2 表示在基线（YOLOv5s）上使用 RepGFPN 替换 FPN 结构，由实验结果可知，mAP 得到了 1.6%的提升，说明 RepGFPN 有加强网络多尺度特征提取能力的作用，能有效融合不同尺度特性；Experiment3 表示在 Experiment1 的基础上再对不同粒度的特征进行聚合，与 Experiment1 方法相比，mAP 增加了 0.8%，使目标检测性能得到进一步提升。

表 4-3 在 VOC 2007 测试集上的消融实验结果(%)

Method	SK	RepGFPN	mAP
YOLOv5s	×	×	85.27
Experience1	√	×	86.28
Experience2	×	√	86.90
Experience3	√	√	87.09

表 4-4 不同类别的检测精度对比(%)

Category	YOLOv5s	Experiment1	Experiment2	Experiment3
aeroplane	93.55	93.48	94.86	94.96
bicycle	92.56	92.29	92.64	93.90
bird	85.38	85.90	86.37	86.20
boat	76.53	81.12	79.90	80.64
bottle	82.87	82.08	81.32	83.99
bus	92.01	92.63	93.25	93.76
car	94.18	95.14	95.03	95.23

cat	80.21	81.57	84.71	83.20
chair	75.43	76.06	75.58	76.38
cow	89.42	90.69	90.81	90.64
dining table	79.71	81.61	81.87	84.73
dog	81.80	82.56	84.88	83.77
horse	89.13	91.64	91.94	92.03
motorbike	92.22	92.10	93.02	92.35
person	91.66	92.51	92.68	92.50
potted plant	64.21	65.92	67.73	66.93
sheep	87.00	86.77	88.83	87.35
sofa	82.51	83.16	82.33	82.91
train	85.97	88.90	89.90	90.55
tvmonitor	89.06	89.37	90.27	89.67
mAP	85.27	86.28	86.90	87.09

根据表 4-4 中不同类别的检测精度数据分析结果显示：引入 SK 注意力后，基线网络在处理多尺度问题时表现出更强的能力。相比于未引入 SK 注意力的基线网络，大多数物体类别的检测精度有所提升，例如，船、公交车和火车等大物体类别的平均精度（AP）分别增加了 4.59%、0.62%和 2.93%；餐桌和花盆等中型物体类别的 AP 分别提高了 1.9%；而鸟和猫等小物体类别的 AP 分别提高了 0.52%和 1.36%。SK 注意力模块的引入有助于网络学习选择融合不同感受野特征图信息，以提高检测性能。在基于 Experiment1 的基础上加入 RepGFPN 后，针对特征融合网络在高层语义信息和低层空间信息融合方面存在不充分的问题，RepGFPN 通过跳跃连接和重参数化有效融合不同层次的特征信息。实验结果表明 RepGFPN 有效地提升了网络的多尺度特征表达能力。例如在 bus（公交车）、dining table（餐桌）、bottle（杯子）这些类别 AP 分别又有了 1.13%、3.12%、1.91%的提升，使得改良后的模型在 PASCAL VOC 数据集上的 mAP 提高至 87.09%。

4.3.3 与其他方法进行比较

为了验证本章算法的有效性，对比实验在公共数据集 PASCAL VOC 上进行，将所提算法与现有的目标检测先进算法进行比较，主要算法包括 Faster R-CNN[27]、SSD[35]、CenterNet[43]、FCOS[45]、YOLOv3[36]、YOLOv5[38]、YOLOv8、文献[69]、文献[70]等，具体实验数据如表 4-5 所示。

表 4-5 在 VOC2007 测试集和其他方法对比实验结果(%)

Method	**Backbone**	**mAP**
Faster R-CNN[27]	ResNet-101	76.4
SSD[35]	VGG-16	74.3
CenterNet[43]	ResNet-50	80.1
FCOS[45]	ResNet-50	78.7
YOLOv3[36]	DarNet53	80.3
YOLOv5[38]	CSPDarNet53	85.3
YOLOv8	CSPDarNet53	86.4
文献[69]	CSPDarNet53	71.0
文献[70]	CSPDarNet53	84.6
第三章算法	ResNet-50	83.8
Ours	**CSPDarNet53**	**87.1**

本章提出的基于多感受野注意力和特征融合的目标检测算法与基线（YOLOv5s）算法相比，在 VOC2007 测试集上 mAP 增加了 1.8%。由于 YOLOv5s 算法采用的网络结构多尺度特征提取能力不足，缺乏有效的感受野以及高层语义信息和低层空间信息的充分交互，导致多尺度目标识别准确率低，本章通过引入 SK 和 RepGFPN 等两方面改进原始 YOLOv5s 网络，挖掘深层特征的多感受野信息，加大深层网络对不同尺度下有用信息的学习力度，并将输出的多尺度特征增强表达图和多维度的空间位置信息进行融合，增强网络对多尺度信息的提取能力，从而提升目标检测性能。本章算法与 Faster R-CNN 算法相比，mAP 增加了 10.7%，Faster R-CNN 算法使用共享的全局卷积特征图来生成候选区域，在全局卷积中每个像素点获得的感受野范围相同，没有考虑到不同尺度目标对不同大小感受野的需求，本章算法使用 SK 注意力在深层网络中挖掘多感受野，学习不同尺度感受野下的关键信息，实现了网络感受野的拓展与丰富；与 SSD 算法相比，mAP 增加了 12.8%，SSD 算法使用不同分辨率的特征图直接预测目标，每一层特征图的独立处理缺乏本章对其他尺度的上下文信息的学习。与 CenterNet 算法相比，mAP 增加了 7%，CenterNet 算法使用单一分辨率的输出预测所有尺寸的目标，缺乏对不同尺度特征的学习，本章算法使用重新参数化的广义特征金字塔融合网络对不同层级的特征信息进行合并，通过跳级连接和重参数化的方式来提取不同尺度下的特征，实现不同层级特征信息的充分交换。与 YOLOv8 算法相比，mAP 增加了 0.7%，YOLOv8 算法的特征融合

网络只关注到相邻层的共享特征，本章算法通过采用重参数思想和层聚合连接结构的FB模块来合并相邻上下层以及同一层级的不同尺度特征，以实现相邻层以及上下层的共享特征的学习，进一步加强了网络的多尺度特征提取能力。与在同一基线(YOLOv5s)算法上改进的文献相比，mAP 增加了 0.5%，文献[70]通过在主干网络中构建多条不同尺度的并行卷积支路来融合多尺度的特征信息，提高了网络的多尺度目标检测能力，但缺乏本章对不同尺度特征图中关键信息的关注。与第三章算法相比，mAP 增加了 3.3%，第三章算法通过提取上下文信息和融合多级特征的方式来丰富多尺度特征和减少位置信息损失，但仅依靠单一尺度特征输出结构处理图像特征，忽略不同尺度目标特征的完整表达对输出分辨率大小的需求，本章算法利用特征金字塔和特征融合的思想输出多个检测头在不同尺度的特征图上进行预测，有效地缓解由固定尺度特征引起的重要目标信息丢失的问题，同时通过多感受野注意力模块学习图像中不同尺度特征的关联性，进一步加强了网络的多尺度特征提取能力。

通过实验结果表明本章算法对多感受野特征的学习能力更好，更能有效融合高层语义信息和低层空间信息，更能应对多尺度目标识别任务，进一步证明了所提算法的有效性。

4.3.4 主观评价结果

为了验证所提的方法的先进性，如图 4-5 所示，将基线和所提方法在 PASCAL VOC 数据集上获得的识别结果进行可视化。其中，第一行图像为基线算法检测结果，第二行图像为改进算法的检测结果。

识别结果显示，本章方法在实际应用场景中的表现优于基线算法。例如，从第一列和第二列识别对比图可以看出，图中不同大小目标的置信度得分有了不同幅度的提升，锚框的定位也更准确，得益于本章方法在基线算法的基础上引入SK注意力，可以自适应地调整网络的感受野，能满足不同尺度目标对感受野的需求，从而加强了网络对多尺度目标位置的度量能力。第三列识别对比图可以看出，原算法出现了将牛(cow)误检为羊(sheep)，本章算法通过采用 RepGFPN 模块作为基线算法的特征融合结构对不同层次特征进行融合，加强了网络的特征提取能力，有效降低了误检。在第四列的识别对比图可以看出，图中的目标尺度差异大，基线算法检测时出现了漏检的情况，而本章算法通过引

入 SK 注意力模块和 RepGFPN 模块，在扩大了网络感受野的同时加大了对多尺度特征的学习力度，从而能准确检测到飞机(aeroplane)这类大物体。因此，本章提出的方法可以更好地完成目标检测任务。

（a）Baseline

（b）Ours

图 4-5 可视化检测结果对比图

4.4 本章小结

本章提出了一种基于多感受野注意力和特征融合的目标检测改进算法。该方法在骨干特征提取网络之后嵌入多感受野注意力模块，利用多路径卷积学习不同尺度特征的关联性，弱化相关目标干扰，实现了对多尺度特征的有效提取。同时利用重新参数化特征金字塔网络对深层特征和不同层次特征进行融合，进一步提高了网络对多尺度特征的捕获能力。最终在 PASCAL VOC 公开数据集上进行实验，mAP 达到了 87.1%，约比基线算法 YOLOv5s 增加了 1.8%，相较于经典算法 SSD、CenterNet、YOLOv8 分别增加了 12.8%、7.0%、0.7%。实验结果表明，本章方法在一定程度上缓解了相关目标干扰和关键信息丢失引起了检测能力不佳问题，有效地提升了多类别目标的检测性能，能够应对智能家居等领域中目标的多样化和尺度差异的挑战。但是也存在一些局限性，例如小目标的识别准确度低。

第 5 章 融合多尺度特征的小目标检测算法

5.1 引言

在计算机视觉领域内，小目标的检测一直是一项颇具挑战的任务，主要由于小目标通常具有低分辨率、复杂的背景等难以处理的特性，传统方法往往难以准确地识别和定位。在工地安全管理领域，安全帽佩戴检测是一项重要的应用，对于工人而言，安全帽可以保护工人的头部免受外部伤害，是工人的一种安全保障，因此，通过监管安全帽佩戴行为对于减少工地意外具有重要的现实意义。然而，在摄像头视角下拍到的图像中安全帽通常占据较小尺寸和展现出低对比度，这使得目前的目标检测模型在提取安全帽等小目标特征以及精确定位其位置方面面临重大挑战。

论文第三章和第四章都利用特征融合策略学习更多的目标特征完成目标检测算法中的特征提取任务，并通过结合多种损失来优化网络参数，同时还采用感受野扩展策略，以学习上下文信息辅助目标判断，这样的设计方式明显提升了基线算法的性能，然而，在复杂多变且具有挑战性的目标检测应用场景中，当出现分辨率较低的小目标时，这种特征融合方法没有关注到来自网络更浅层的丰富的细节特征，并且模型输出的检测层分辨率不足会直接引起小目标细节信息的损失，导致针对小目标表征的判别力不足，使得第三章和第四章对于小目标的应用检测仍存在局限性。

本章着重考虑挖掘更浅层的特性，以达到充分提取细节信息的目的，此外，考虑到监控场景下对检测速度的需求以及实际应用中轻量化模型更易于部署，于是，采用模型小、检测速度快的 YOLOv5s 作为基线模型，从关注底层特性和优化损失函数的角度出发，提出一种融合多尺度特征的小目标检测改进算法。首先，在 YOLOv5s 的颈部特征融合结构中添加小目标检测层(Small_Layer)，通过使用拼接层将深层特征和更浅层的特征进行融合以输出四种尺度的特征层，其中输出的高分辨率的检测头具有丰富的小目标细节特征，用以检测小目标，从而实现对小目标的准确检测，并将融合后的多尺度特征信息传递到更深层的网络，以学习丰富的上下文特征和细节特征辅助大目标判断。其次，联合 K-

means 聚类算法重新设计尺度更精细的锚框参数，提高先验包围盒与不同尺度目标的匹配度。最后，采用 NWD 距离度量方法来优化损失的计算方式，通过联合 NWD 和 CIoU 损失对网络捕获的特征进行联合监督，以弱化 CIoU 损失计算小物体位置偏差的敏感度，从而提高模型的定位能力。

5.2 算法流程

5.2.1 网络结构设计

本章提出一种基于 CSPDarkNet53 骨干网络的网络框架，用于安全帽这种小目标的检测算法。通过构建 Small_Layer、引入 K-means 聚类算法和 NWD 损失三方面对原始的 YOLOv5s 模型进行改进，以提高模型对目标识别的准确性。改进后的网络结构由主干特征提取网络（CSPDarkNet53）、特征融合网络（FPN）、小目标检测层（Small_Layer）和检测头（Head）四个部分组成。主干特征提取网络负责学习输入图像的特征，在输出中包含不同层次的信息。随后，特征融合网络将主干网络输出的深层特征进行融合，并通过小目标检测层融合更浅层的特征，最终生成四个检测头，以学习图像的四种尺度特征。最后，采用 K-means 聚类算法重新设计预设锚框的大小和长宽比例，实现先验包围盒与目标匹配度的提高，并联合 NWD 和 CIoU 定位损失对网络捕获的特征进行联合监督，弱化网络对小尺度目标的位置偏差的敏感度，从而实现对小目标的准确检测。

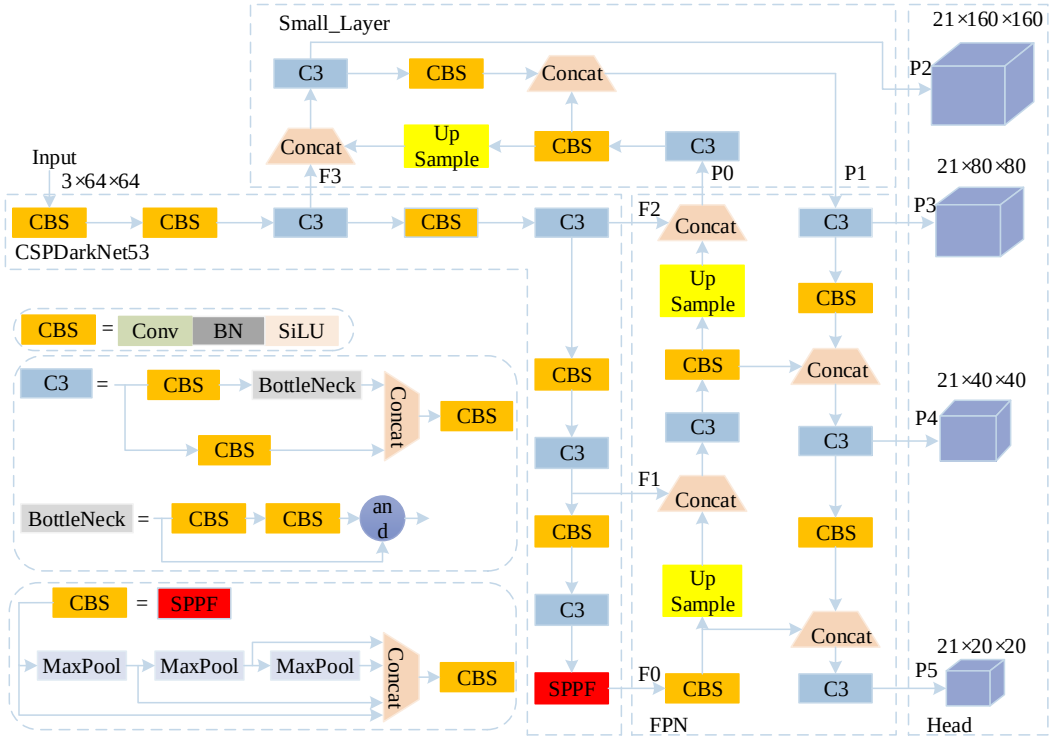

图 5-1 整体网络框架

5.2.2 小目标检测层 small_Layer

在 YOLOv5s 主干特征提取网络经过多次下采样处理后，会丢失大量低级特征，导致高层特征难以准确捕捉小目标的细节信息。为了解决这个问题，提出了一种改进方法，即设计了一个小目标检测层（Small_Layer），通过特征融合和通道维度调整的操作融合更浅层的特征，以学习四种尺度特征，从而帮助小目标检测模型更有效地捕获目标的特征和上下文信息。结构如图 5-1 所示，其中，由虚线条框出的 Small_Layer 模块为小目标检测层。

由图 5-1 可知，首先，通过特征融合网络在主干特征提取网络输出的特征图中对多个尺度的特征图进行融合，生成多尺度特征图 P0。接着，为了让网络学习到低级特征，采用 C3 和 CBS 结构对特征图 P0 进行增强特征提取，并经过上采样处理，将增强后的特征图与更浅层的特征图进行融合，形成包含低级特征信息的多尺度特征图 F4。接下来，使用普通卷积来调整特征图的通道维度，以输出具有高分辨率的检测头。同时，继续进行下采样操作，将低级特征传递给更深层的网络，以进一步提高网络对不同尺度特征的学习和利用，从而提高安全帽佩戴检测的识别精度。

5.2.3 K-means 聚类算法

YOLOv5s 的原始网络只有三个检测头，如图 5-1 所示，即 P3、P4 和 P5，三个检测头分别对应三组预设锚框。本章对原始的 YOLOv5s 进行改进，通过小目标检测层来融合浅层特征后，最终输出四个检测头，分别为P2、P3、P4和P5，因此，三组预设的锚框尺度无法匹配四个检测头。此外，在基于 YOLOv5s 的目标检测中，三组锚框是针对 coco 数据集[71]中的目标尺度来设计的，并不适用于安全帽佩戴检测数据集(Safety Helmet Wearing Dataset, SHWD)。本章为了更好地适应安全帽佩戴检测数据集，重新设计了四组适合该数据集的锚框，以提高先验包围盒与目标的匹配度。为此，利用 K-means 聚类算法对训练集中的目标边界框进行聚类，以更好地适应 SHWD 数据集的目标特征尺度。在对边界框聚类时，以边界框的宽度和高度作为特征，对边界框的宽度和高度进行归一化，公式如（5-1）（5-2）所示。

$$w = \frac{w_{\text{box}}}{w_{\text{img}}} \tag{5-1}$$

$$h = \frac{h_{\text{box}}}{h_{\text{img}}} \tag{5-2}$$

式中，w_{box}，h_{box} 分别为边界框的宽度和高度，h_{img}，w_{img} 为输入图像的宽度和高度。对边界框 K-means 聚类的步骤是：

步骤 1：随机选择 K 个边界框作为预设锚框；

步骤 2：利用 IOU 度量将各个方框分配给其最近的锚框；

步骤 3：算出各个聚类中全部框的高度和宽度的平均值且更新锚点；

步骤 4：重复以上步骤，直至达到最大迭代次数，或锚框的尺寸不再改变。

在 YOLOv5s 的输出处有四个检测头。K-means 聚类后的锚框的大小如表 5-1 所示。

表 5-1 调整后的锚框

检测层	聚类后
160×160	[8,18, 11,24, 16,32]
80×80	[24,43, 35,59, 47,83]
40×40	[69,110, 86,159, 117,238]
20×20	[157,166, 192,317, 345,439]

5.2.4 损失函数的改进

YOLOv5s 的损失函数由分类损失、定位损失和置信度三种损失组成，其中，通过采用 CIoU[72]损失作为定位损失来测评真实框(Ground Truth，GT)和预测框(Predicted Truth ,PT)之间的距离。CIoU 损失同时衡量了真实框和预测框二者间的重叠面积、质心距离和宽高比，计算公式如式（5-3）、（5-4）、（5-5）、（5-6）所示。

$$L_{\mathrm{CIoU}} = 1 - \mathrm{IoU} + \frac{\rho^2(b, b^{gt})}{\mathrm{c}^2} + \alpha v \tag{5-3}$$

$$v = \frac{4}{\pi^2}(\arctan\frac{w^{gt}}{h^{gt}} - \arctan\frac{w}{h})^2 \tag{5-4}$$

$$\alpha = \frac{v}{(1-\mathrm{IoU}) + v} \tag{5-5}$$

$$\mathrm{IoU} = \frac{\mathrm{PT} \cap \mathrm{GT}}{\mathrm{PT} \cup \mathrm{GT}} \tag{5-6}$$

其中，b 和 b^{gt} 表示 PT 和 GT 的中心点，$\rho^2(b, b^{gt})$ 表示 PT 和 GT 中心点之间的欧氏距离，c 表示包含 GT 和 PT 的最小长方形的对角线长度，v 表示 PT 和 GT 的长宽比的相似度，w^{gt}, h^{gt} 和 w，h 分别表示 PT 和 GT 的宽度和高度，α 为平衡系数。公式（5-3）和（5-5）中，1-IoU 为 IoU 损失，IoU 损失旨在消除训练和测试之间的性能差距。但是，在以下两种情况下，IoU 损失不能为优化网络提供梯度：（1）预测框和真实框之间没有重叠。（2）预测框完全包含真实框，反之亦然。这两种情况对于小物体来说很常见。具体来说，一方面，预测框中几个像素的偏差会导致 PT 和 GT 之间没有重叠，另一方面，小物体容易被错误预测，导致 PT 完全包含 GT(或者 GT 完全包含 PT)。因此，IoU 损失不适用于小物体检测。CIoU 损失是基于 IoU 损失提出的，对小物体的位置偏差同样很敏感。为了解决上述问题，在回归损失函数中，添加了 NWD 损失来弥补 CIoU 损失对于小目标检测的缺点，但仍然保留 CIoU 损失用于检测中大型目标，以实现对不同尺度目标的均衡评估。

NWD[73]是一种新的小目标检测评估方法，具有尺度不敏感性，特别适用于测量小目标之间的相似度。该距离度量方法将边界框建模为二维高斯分布，通过相应的高斯分布计算预测目标与真实目标之间的相似度。对于已检测到的目

标，无论它们是否重叠，都可利用分布相似度来评估预测目标与真实目标之间的距离。其损失函数公式如（5-7）、（5-8）、（5-9）所示。

$$L_{\mathrm{NWD}} = 1 - NWD(N_{\mathrm{a}}, N_{\mathrm{b}}) \tag{5-7}$$

$$NWD(N_{\mathrm{a}}, N_{\mathrm{b}}) = \exp(-\frac{\sqrt{W_2^2(N_{\mathrm{a}}, N_{\mathrm{b}})}}{\mathrm{C}}) \tag{5-8}$$

$$W_2^2(N_{\mathrm{a}}, N_{\mathrm{b}}) = \left\| [cx_{\mathrm{a}}, cy_{\mathrm{a}}, \frac{w_{\mathrm{a}}}{2}, \frac{h_{\mathrm{a}}}{2}]^T, [cx_{\mathrm{b}}, cy_{\mathrm{b}}, \frac{w_{\mathrm{b}}}{2}. \frac{h_{\mathrm{b}}}{2}]^T \right\|^2 \tag{5-9}$$

其中，C 是一个数据集密切相关的常数，$W_2^2(N_{\mathrm{a}}, N_{\mathrm{b}})$是一个距离度量，$N_{\mathrm{a}}$和$N_{\mathrm{b}}$是由$A = (\mathrm{cx}_{\mathrm{a}}, \mathrm{cy}_{\mathrm{a}}, w_{\mathrm{a}}, h_{\mathrm{a}})$和$B = (\mathrm{cx}_{\mathrm{b}}, \mathrm{cy}_{\mathrm{b}}, w_{\mathrm{b}}, h_{\mathrm{b}})$建模的高斯分布。

改进后的 YOLOv5 的总损失函数公式如式（5-10）、（5-11）所示。

$$L_{\mathrm{total}} = L_{\mathrm{box}} + L_{\mathrm{cls}} + L_{\mathrm{obj}} \tag{5-10}$$

$$L_{\mathrm{box}} = \alpha L_{\mathrm{CIoU}} + (1-\alpha) L_{\mathrm{NWD}} \tag{5-11}$$

其中，分类损失L_{cls}和置信度损失L_{obj}使用二进制交叉熵损失函数（binary cross entropy loss，BCE Loss）[74]计算，α和 1-α用于调控两部分损失贡献的比例系数。

5.3 实验设置与结果分析

5.3.1 实验设置和数据集

本章实验环境配置表 5-2 所示，操作系统为 Windows10，CPU 为 12 核 Intel(R) Xeon(R) Platinum 8255C CPU @ 2.50GHz，内存 43G，显卡 Nvidia GeForce RTX 2080Ti(11GB)，CUDA版本为11.1，深度学习框架为pytorch1.9.0。

表 5-2 实验环境配置

实验配置	型号或参数
系统	Windows10
CPU	12 vCPU Intel(R) Xeon(R) Platinum 8255C CPU @ 2.50GHz
GPU	Nvidia GeForce RTX 2080Ti（11G）
CUDA	11.1
深度学习框架	Pytorch1.9.0
算法语言	Python

实验参数设置同上一章节，实验过程是利用 SHWD 数据集对模型进行评估,该数据集包含 7571 张图像，总共包含 7459 个佩戴安全帽的对象（hat）和 100036 个未佩戴安全帽的对象（person）。这些图像是在真实的建筑工地上从各种角度拍摄的，适用于多类对象(佩戴安全帽的对象和未佩戴安全帽的对象)检测的主要数据集。在训练前，将数据集中的图片按 9:1 的比例分为训练集和测试集。其数据集组成如下表 5-3 所示。

表 5-3 SHWD 数据集训练和测试数据统计

Category	Trainval		Test	
	Images	Objects	Images	Objects
hat	6813	6736	758	723
person		90193		9843
Total	6813	96929	758	10566

5.3.2 消融实验

本章中，在公共数据集 SHWD 数据集下进行消融实验，来评估所提网络和模块的有效性，如表 5-4 所示，。

表 5-4 SHWD 数据集消融实验比较(%)

Method	Small_Layer	k-means	NWD	mAP
YOLOv5s	×	×	×	92.8
Experience1	✓	×	×	93.9
Experience2	✓	✓	×	94.2
Experience3	×	×	✓	94.7
Experience4	✓	×	✓	95.3
Experience5	✓	✓	✓	**95.5**

其中，Small_Layer 表示通过搭建小目标检测层对网络的低层特性进行学习，k-means 表示对模型使用聚类算法进行锚框参数的调整，NWD 表示对网络使用归一化高斯距离分布损失进行监督学习。其中，Experience1 表示在基线网络(YOLOv5s)的基础上添加 Small_Layer，由实验结果可知，采用设计的 Small_Layer 与基线相比性能得到明显提升，mAP 增加了 1.4%，说明 Small_Layer 有细化和丰富多尺度特征的作用，能有效学习低层特性。Experience2 表示在 Experiment1 的基础上再使用 k-means 算法对网络的锚框参数

进行优化，mAP 增加了 0.3%，说明 k-means 算法有提高先验锚框与目标匹配度的作用，能有效适应数据集中目标的真实分布。Experiment3 表示在基线网络的基础上添加 NWD 损失，以联合原有的损失对网络进行监督学习，由实验结果可知，与基线算法相比，mAP 得到了 1.9%的提升，说明 NWD 损失有提升模型的目标检测能力的作用，能有效弱化小尺度目标的位置偏差敏感度。Experiment4 表示在 Experiment1 的基础上对网络使用多种损失联合监督，与 Experiment1 方法相比，mAP 增加了 1.4%，Experiment5 表示在 Experiment2 的基础上再对网络使用多种损失联合监督，与 Experiment2 方法相比，mAP 增加了 1.3%，使安全帽佩戴小目标检测性能得到进一步提升。本章先采用 NWD 代替 CIoU 作为回归损失，检测效果得到了大幅度提升，又考虑到 SHWD 数据集中虽然多小目标，但也存在一定比例的中大型目标，直接采用 NWD 代替 CIoU 检测效果可能不能达到最优。因此，本章保留原有的 CIoU 损失再添加 NWD 损失，以保障不丢失中大型目标的精度的同时，平衡 CIoU 损失在评估小目标时的不足，并通过调整它们之间的比例关系来提高本章模型对安全帽佩戴目标检测的鲁棒性。

如表 5-5 所示，为探索本章中 CIoU 和 NWD 两种损失的权重系数对网络性能的影响，在 SHWD 数据集下设计了一系列对比实验。设计的定位损失形式为 $L_{box}=\alpha L_{CIoU}+(1-\alpha)L_{NWD}$，其中，$\alpha$ 和 1-α 分别表示 CIoU 和 NWD 两种损失的权重系数，通过调整二者的比例系数以求得最优比例配。实验结果表明，CIoU 损失和NWD损失系数分别设置为0.8和0.2时，检测效果达到最优。具体点来说，本章提出了将 CIoU 和 NWD 两种回归损失联合在一起的目标检测改进方法，该方法旨在提高安全帽佩戴目标检测的鲁棒性，CIoU 损失主要用于度量中大型目标之间的相似度，NWD 对目标的尺度不敏感，用于测量小目标之间的相似度，通过调整平衡系数 α 为 0.8，使得在安全帽佩戴检测方面的表现得到了最优化，最终的平均mAP相对于基线提升了1.9%，说明了本章提出的结合CIoU和NWD两种损失的方法的有效性。

表 5-5 NWD 损失函数实验评估结果及对比(%)

Method	***α***	**1-*α***	**AP**		**mAP**
			hat	**person**	
Experience1	1	0	92.4	93.1	92.8

Experience2	**0.8**	**0.2**	**95.2**	**94.2**	**94.7**
Experience3	0.5	0.5	95.3	93.7	94.5
Experience4	0.2	0.8	95.2	93.8	94.5
Experience5	0	1	95.3	93.6	94.4

5.3.3 与其他方法进行比较

为了验证本章算法的有效性，对比实验基于公共数据集 SHWD 数据集，将所提算法与现有的目标检测先进算法进行比较，主要包括 SSD[35]、YOLOv3[36]、CenterNet[43]、YOLOv4[37]、YOLOv5[38]、YOLOv6[75]、YOLOv7[76]、YOLOv8、文献[77]、文献[78]、文献[79]等。实验数据如表 5-6 所示，其中“—”表示原论文中没有报告的结果。

表 5-6 不同算法在 SHWD 上的 mAP 对比(%)

Method	hat	person	mAP	FPS	Model Size/MB
SSD[35]	83.1	37.9	60.5	113.1	100
YOLOv3[36]	92.1	94.3	93.2	135	118
CenterNet[43]	91.8	87.9	89.9	79	125
YOLOv4[37]	89.8	82.1	85.9	51	256
YOLOv5[38]	92.4	93.1	92.8	277	14
YOLOv6[75]	86.8	91.8	89.3	54	39
YOLOv7[76]	91.4	94.3	92.8	128	71
YOLOv8	94.1	94.7	94.4	400	22
文献[77]	—	—	89.6	32	—
文献[78]	85.6	88.9	87.2	49.2	—
文献[79]	—	—	91.0	137	23
Ours	**95.5**	**95.5**	**95.5**	**179**	**16**

本章提出的融合多尺度特征的小目标检测改进算法与基线(YOLOv5s)算法相比，在 SHWD 测试集上 mAP 增加了 2.7%。由于 YOLOv5s 算法的预设锚框尺度不合适，同时缺乏对底层特性和上下文信息的关注，导致对工地场景中监控镜头下是否佩戴安全帽的人的识别能力低，本章通过构建 Small_Layer、引入 K-means 聚类算法和 NWD 损失等三方面改进原始 YOLOv5s，通过学习目标的浅层信息，再将融合后的特征信息传递到不同维度的网络，以丰富上下文信息和细化小目标特征，同时调整预设锚框的参数和使用多种损失对提取的特征联合监督，从而提升安全帽佩戴检测性能。

本章算法与 SSD 算法相比，mAP 增加了 35%，检测精度获得大幅度提升，SSD 算法使用不同分辨率的特征图直接预测目标，每一层特征图的独立处理缺乏本章对其他层次信息的学习，本章算法在处理图像输入时采用了多尺度融合的策略，又通过搭建小目标检测层将浅层特征传递到不同深度的网络，加强了小目标检测模型捕获目标特征和上下文信息的能力。与 CenterNet 算法相比，mAP 增加了 5.6%，CenterNet 算法利用高斯热力图来标注目标中心点进行定位通过中心点来预测目标位置，对不规则目标边界信息的学习能力不佳，本章算法采用的直接预测目标边界框的方法，通过引入中心点距离和长宽比调整因子，加强了网络对目标边界框的形状变化信息的学习，又联合 NWD 损失平衡网络缓解小尺度目标位置偏差的敏感度，进一步加强了网络对小目标的识别能力。与 YOLOv8 算法相比，mAP 增加了 1.1%，YOLOv8 算法利用金字塔特征融合网络对不同层次特征进行融合，只关注到图像的三种尺度特征，没有考虑对更丰富的维度信息进行学习，本章算法通过联合小目标检测层和基线网络 YOLOv5s 的检测层对目标特征进行提取，实现了网络对图像四种尺度特征的学习，又联合 K-means 聚类算法重新生成尺度更精细的锚框参数，在丰富了特征的同时提高了先验包围盒与目标匹配度。与在同一基线(YOLOv5s)算法上改进的文献[79]。相比，mAP 增加了 4.5%，文献[79]采用双向聚合网络对不同层级特征进行信息交互，以提升多尺度特征的学习能力，但缺乏本章对更浅层的特征信息的关注。

通过实验结果表明本章算法对是否佩戴安全帽的人这类小目标的特征提取效果更好，能更准确地度量感兴趣目标的大小和位置，更能应对工地场景中安全帽佩戴检测的任务，进一步证明了所提算法的有效性。

5.3.4 可视化结果

为了验证所提的方法的先进性，如图 5-2 所示，将基线和所提方法在 SHWD 数据集上获得的识别结果进行可视化。其中，第一行图像为基线算法检测结果，第二行图像为改进算法的检测结果。识别到带有红色框的目标图像表示佩戴了安全帽的人，识别到带有粉色框的目标图像表示未佩戴安全帽的人。

（a）Baseline

（b）Ours

图 5-2 可视化识别结果比较

识别结果显示，本章方法在安全帽佩戴检测场景中的表现优于基线算法。例如，从第一列的识别对比图可以看出，图中小目标的置信度得分有了不同幅度的提升，锚框的定位也更准确，得益于本章方法在基线算法的基础上引入 NWD 损失，有弱化小尺度目标位置偏移敏感度的作用，从而加强了网络对小目标位置的度量能力。从第二列的识别对比图可以看出，原算法出现了将戴了安全帽的人(hat)误检为未戴安全帽的人(person)，本章算法对基线算法改进后，加强了网络的多尺度特征提取能力，有效降低了误检。在第三列和第四列的安全帽佩戴检测场景中，基线算法存在漏检，而本章算法能准确检测到远处的戴安全帽的人(hat)。因此，本章提出的方法可以更好地适应安全帽。

5.4 本章小结

本章从充分提取小目标信息和优化损失计算方法的角度出发，针对安全帽佩戴检测任务提出了一种联合小目标检测层的多尺度安全帽佩戴检测算法。网络通过在颈部特征融合网络之后添加小目标检测层，利用拼接层提取底层特性并将浅层信息传递到不同深度的网络，加强了网络对四种尺度特征的学习。同时联合 NWD 损失来评估不同尺度目标的分布相似度，减小网络对不同尺度目标位置偏差敏感度的差异，提高了网络对多尺度目标的定位能力。最终在 SHWD 公开数据集上进行实验，mAP 达到了 95.5%，约比基线算法 YOLOv5s 增加了 1.8%，相较于经典算法 SSD、CenterNet、YOLOv8 分别增加了 12.8%、7.0%、0.7%。实验结果表明，本章方法在一定程度上缓解了小目标信息丢失和

尺度敏感度差异大引起了安全帽佩戴检测能力不佳问题，在工地安全管理领域的应用中更具优越性。

第 6 章 总结与展望

6.1 工作总结

随着人工智能的兴起，深度学习技术发展迅速，计算机的分析处理能力不断增强。近年来，目标检测算法因其突破性进展和广泛应用备受关注。目标检测作为计算机视觉中的基础任务之一，是进行目标跟踪、动作识别和行为表达的必要前提。论文主要围绕如何有效提取不同尺度目标特征的问题展开研究，从多尺度特征图预测、特征融合、注意力机制、锚框优化和损失函数等方面进行研究和分析。现对论文主要研究内容总结如下：

（1）本文深入探讨了目标检测领域的众多研究成果，对当前目标检测技术的研究现状及其理论基础进行了系统的梳理与分析。同时，对目前目标检测技术研究中所面临的关键难题进行了详细的梳理与总结。

（2）提出了上下文特征增强和特征融合的目标检测改进方法，采用多感受野技术和信息融合策略。设计了自适应上下文提取模块提取深层网络不同感受野下的特征信息，加强网络的上下文特征提取能力，其中使用 ACON-C 激活函数改善网络的表示能力和泛化能力，进一步提高了目标的检测能力；此外，为减少浅层位置信息损失，采用特征融合策略，对不同层次特征进行合并，加强了对浅层位置信息的学习力度。解决了 CenterNet 算法在多尺度目标识别中由感受野受限以及浅层位置信息缺失造成的表征能力不足的问题，提高了模型的多尺度目标识别能力和定位准确度。

（3）提出了联合多感受野注意力和特征融合的目标检测改进方法。为缓解不同感受野下的特征权重分配不当引起的目标间相互干扰问题，在主干特征提取网络之后，采用多感受野注意力模块自适应地对特征进行加权，通过学习多分支特征的关联性形成适当的特征权重分配，弱化了不同尺度特征间的相互干扰；此外，采用重新参数化的广义特征金字塔融合网络挖掘上下层以及相邻层的特征信息，减少了重要信息的丢失。解决了 YOLOv5s 算法在多尺度目标识别中由不同感受野特征间相互干扰引起的精度下降问题。

（4）提出了融合多尺度特征的小目标检测改进方法，采用信息融合和优化

损失函数的思想。设计了小目标检测层从浅层网络提取目标细节信息，丰富小目标特征，辅助网络提高对小目标的判别能力，其中使用 k-means 聚类算法重新生成锚框参数，提高了锚框匹配度；此外，为提高网络对小目标表征相似度的准确度，引入了 NWD 损失评估小目标的真实框和预测框之间的分布相似度，多种损失的联合监督加强了模型对小目标位置的度量能力。解决了 YOLOv5s 算法在检测安全帽佩戴行为时由输出特征分辨率欠缺引起的小目标表征能力不足和识别准确率降低的问题。

6.2 工作展望

目标检测是指识别图像中的目标并标出其位置，通过分类输出图片中感兴趣目标的类别信息，同时通过矩形框框出输出目标的位置。虽然论文从目标检测任务策略的角度出发，提出了多种加强多尺度特征提取的方法，以更准确地检测物体。这些方法在一定程度上提高了检测算法的准确性，并优化了目标检测的效果，但是对于密集目标的场景仍然存在漏检和误检的情况，因此未来会将重心集中在以下几个方面：

（1）注意力机制方面。本文在第三章模型中仅在特征融合模块加入了注意力机制，在第四章模型中仅在骨干网络之后加入了注意力模块，未来将尝试在骨干网络中运用多样的注意力机制，以提高模型的多尺度特征提取能力。

（2）数据增强方面。本文没有尝试新的数据增强处理，未来将尝试使用一些最新的数据增强方法，与原论文作对比。

参考文献

[1] Zhou D L, Wang X D. Robust infrared small target detection using a novel four-leaf model[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2024, 17: 1462-1469.

[2] Cui C H, Wang R G, Wang Y Y, et al. Research on optical remote sensing image target detection technique based on DCH-YOLOv7 algorithm[J]. IEEE Access, 2024, 12: 34741-34751.

[3] Kononov A A, Ka M H. Rapid adaptive matched filter for detecting radar targets with unknown velocity[J]. IEEE Access, 2024, 25411-25428.

[4] Zhou X, Xu X, Liang W, et al. Intelligent small object detection for digital twin in smart manufacturing with industrial cyber-physical systems[J]. IEEE Transactions on Industrial Informatics, 2021, 18(2): 1377-1386.

[5] Leng J, Ren Y, Jiang W, et al. Realize your surroundings: Exploiting context information for small object detection[J]. Neurocomputing, 2021, 433: 287-299.

[6] Chang C I. Target-to-anomaly conversion for hyperspectral anomaly detection [J]. IEEE Transactions on Geoscience and Remote Sensing, 2022, 60: 1-28.

[7] Zhang Y D, Gorriz J M, Nayak D R. Optimization algorithms and machine learning Techniques in medical image analysis[J]. Mathematical Biosciences and Engineering, 2023, 20(3): 5917-5920.

[8] Ghaderi O, Abedini M. Evaluation of the airport runway flexible pavement macro-texture using digital image processing technique (DIPT)[J]. International Journal of Pavement Engeering, 2022, 23(13): 4587-4599.

[9] 窦慧, 张凌茗, 韩峰等. 卷积神经网络的可解释性研究综述[J]. 软件学报, 2024, 35(1): 159-184.

[10] Gong Y H, Liu Y K, Gong Z L, et al. A novel detection method for wheat aging based on the delayed luminescence[J]. Scientific Reports, 2024,14(1): 1134.

[11] Zhao C, Chang X K, Xie T, et al. Unsupervised anomaly detection based method of risk evaluation for road traffic accident[J]. Applied Intelligence, 2023, 53(1):

369-384.

[12] Cai J J, Wang Q W, Wang C, et al. Research on cell detection method for microfluidic single cell dispensing[J]. Mathematical Biosciences and Engineering, 2023, 20(2): 3970-3982.

[13] Song R J, Wang Z M. RBFPDet: An anchor-free helmet wearing detection method[J]. Scientific Reports, 2024, 14(1): 1134.

[14] Anusha C, Avadhani P S. Object detection using deep learning [J]. International Journal of Computer, 2018, 182(32): 18-22.

[15] Freund Y, Schapire R. A decision-theoretic generalization of on-line learning and an application to boosting[J]. Journal of Computer and System Sciences, 1997, 55(l): 119-139.

[16] Felzenszwalb P, Girshick R, McAllester D, et al. Object detection with discriminatively trained part-based models[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2009, 32(9): 1627-l645.

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

[18] Lienhart R, Maydt J. An extended set of haar-like features for rapid object detection[C]. Proceedings of the International Conference on Image Processing, Rochester, New York, USA, 2002: 1-8.

[19] Dalal N, Triggs B. Histograms of oriented gradients for human detection[C]. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Diego, CA, USA, 2005: 886-893.

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

[21] Lecun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.

[22] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017, 60(6): 84-90.

[23] Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object

detection and semantic segmentation[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 2014: 580-587.

[24] Uijlings J R R, Van D S, Gevers T, et al. Selective search for object recognition[J]. International Journal of Computer Vision, 2013, 104(2): 154-171.

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

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

[27] Ren S, He K, Girshick R, et al. Faster R-CNN: Towards real-time object detection with 64 region proposal networks[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2017, 39(6): 1137-1149.

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

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

[30] He K, Gkioxari G, Dollár P, et al. Mask R-CNN[C]. Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 2017: 2961-2969.

[31] Pang J, Chen K, Shi J, et al. Libra R-CNN: Towards balanced learning for object detection[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 2019: 821-830.

[32] Liu Z, Mao H, Wu C Y, et al. A convnet for the 2020s[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 2022: 11976-11986.

[33] Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, real-time object detection[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 2016: 779-788.

[34] Redmon J, Farhadi A. YOLO9000: better, faster, stronger[C]. Proceedings of the

IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 2017: 7263-7271.

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

[36] Farhadi A, Redmon J. Yolov3: An incremental improvement[C]. Proceedings of the Computer Vision and Pattern Recognition, Berlin/Heidelberg, Salt Lake City, UT, USA, 2018: 1-6.

[37] Wang C Y, Bochkovskiy A, Liao H Y M. Scaled-yolov4: Scaling cross stage partial network[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 2021: 13029-13038.

[38] Yan B, Fan P, Lei X, et al. A real-time apple targets detection method for picking robot based on improved YOLOv5[J]. Remote Sensing, 2021, 13(9): 1619.

[39] Wang C Y, Yeh I, Liao H M. YOLOv9: Learning What You Want to Learn Using Programmable Gradient Information[J]. arXiv preprint arXiv: 2402.13616v2, 2024.

[40] 刘小波, 肖肖, 王凌等. 基于无锚框的目标检测方法及其在复杂场景下的应用进展[J/OL], 自动化学报，2023, 49(07): 1369-1392.

[41] 李柯泉, 陈燕, 刘佳晨等. 基于深度学习的目标检测算法综述[J]. 计算机工程, 2022, 48(07): 1-12.

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

[43] Duan K, Bai S, Xie L, et al. Centernet: Keypoint triplets for object detection[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Kore, 2019: 6569-6578.

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

[45] Tian Z, Shen C, Chen H, et al. Fcos: Fully convolutional one-stage object detection[C]. Proceedings of the IEEE/CVF International Conference on Computer

Vision, Seoul, Korea, 2019: 9627-9636.

[46] Diamantis D E, Lakovidis D K. Fuzzy pooling [J]. IEEE Transactiona on Fuzzy Systems, 2021, 29(11): 3481-3488.

[47] Wu L, Perin G. On the importance of pooling layer tuning for profiling side-channel analysis[C]. Proceedings of the Applied Cryptography and Network Security Workshops: ACNS 2021 Satellite Workshops, Kamakura, Japan, 2021: 114-132.

[48] Han J, Moraga C. The influence of the sigmoid function parameters on the speed of backpropagation learning[C]. Proceedings of the International Workshop on Artificial Neural Networks, Malaga-Torremolinos, Spain, 1995: 195-201.

[49] Glorot X, Bordes A, Bengio Y, et al. Deep sparse rectifier neural network[C]. Proceedings of the International Conference on Artificial Intelligence and Statistics, Ft. Lauderdale, FL, USA, 2011: 315-323.

[50] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 2016: 770-778.

[51] Wang C Y, Mark Liao H Y, Wu Y H, et al. CSPNet: A new backbone that can enhance learning capability of cnn[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Seattle, WA, USA, 2020: 1571-1580.

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

[53] Wang W, Xie E, Song X, et al. Efficient and accurate arbitrary-shaped text detection with pixel aggregation network[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 2019: 8440-8449.

[54] Ma N N, Zhang X Y, Liu M, et al. Activate or not: learning customized activation[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 2021: 8028-8038.

[55] Yu F, Koltun V, Funkhouser T, et al. Dilated residual networks[C]. Proceedings of

the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 2017: 472-480.

[56] Wang P, Chen P, Yuan Y, et al. Understanding convolution for semantic segmentation[C]. Proceedings of the IEEE Winter Conference on Applications of Computer Vision, Lake Tahoe, NV, USA, 2018: 1451-1460.

[57] Zhang T, Zhang X. Squeeze-and-excitation laplacian pyramid network with dual-polarization feature fusion for ship classification in SAR images[J]. IEEE Geoscience and Remote Sensing Letters, 2022, 19: 1-5.

[58] Everingham M, Eslami S, Van G L, et al. The pascal visual object classes (voc) challenge[J]. International Journal of Computer Vision, 2010, 88(2): 303-338.

[59] Zhu Y, Zhao C, Wang J, et al. Couplenet: coupling global structure with local parts for object detection[C]. Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 2017: 4146-415.

[60] Zhang S, Wen L, Lei Z, et al. Refinedet++: single-shot refinement neural network for object detection[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2021, 31(2): 674-687.

[61] Liu S, Huang D, WangY. Receptive field block net for accurate and fast object detection[C]. Proceedings of the European Conference on Computer Vision, Munich, Germany, 2018: 404-419.

[62] 戴坤, 许立波, 黄世昄等. 融合策略优选和双注意力的单阶段目标检测[J]. 中国图象图形学报, 2022, 27(08): 2430-2443.

[63] 侯志强, 郭浩, 马素刚等. 基于双分支特征融合的无锚框目标检测算法[J]. 电子与信息学报, 2022, 44(06): 2175-2183.

[64] 王启胜, 王凤随, 陈金刚等. 融合自适应注意力机制的 Faster R-CNN 目标检测算法[J]. 激光与光电子学进展, 2022, 59(12): 391-396.

[65] Li X, Wang W H, Hu X L, et al. Selective kernel networks[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 2019: 510-519.

[66] Liu S, Qi L, Qin H, et al. Path aggregation network for instance segmentation[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern

Recognition, Salt Lake City, UT, USA, 2018: 8759-8768.

[67] Jiang X Z, Jiang Y Q, Chen W H, et al. Damo-yolo: A report on real-time object detection design[J]. arXiv preprint arXiv: 2211.15444v4, 2022.

[68] Zhang X D, Zeng H, Guo S, et al. Efficient long-range attention network for image super-resolution[C]. Proceedings of the Lecture Notes in Computer Science, Tel Aviv, Israel, 2022: 649-667.

[69] 薛珊, 安宏宇, 吕琼莹等. 复杂背景下基于 YOLOv7-tiny 的图像目标检测算法[J]. 红外与激光工程, 2024, 53(01): 269-280.

[70] 荆修平, 田莹. 采用长距离依赖和多尺度表达的轻量化车辆检测[J]. 光学精密工程, 2023, 31(06): 950-961.

[71] Tsung Y L, Michael M, Serge B, et al. Microsoft coco: common objects in context[J]. Lecture Notes in Computer Science, 2014, 8693(1): 740-755.

[72] Zheng Z H, Wang P, Liu W, et al. Distance-IoU loss: Faster and better learning for bounding box regression[C]. Proceedings of the AAAI 2020-34th AAAI Conference on Artificial Intelligence, New York, USA, 2020: 12993-13000.

[73] Wang J, Xu C, Yang W, et al. A normalized gaussian wasserstein distance for tiny object detection[J]. arXiv preprint arXiv: 2110. 13389v2, 2021.

[74] Lecun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.

[75] Li C Y, Li L L, Jiang H L, et al. Yolov6: a single-stage object detection framework for industrial applications[J]. arXiv preprint arXiv: 2209. 02976v1, 2022.

[76] Wang C Y, Bochkovskiy A, Liao H Y M. Yolov7: trainable bag-of-freebies sets new state-of-the-art for real-time object detectors[J]. arXiv preprint arXiv: 2207. 02696v1, 2022.

[77] Song R, Wang Z. RBFPDet: An anchor-free helmet wearing detection method[J]. Applied Intelligence, 2023, 53(5): 5013-5028.

[78] Xu Z G, Li Y G, Zhu H L. Safety helmet detection method based on semantic guidance and feature selection fusion[J]. Signal, Image and Video Processing, 2023, 17(17): 3683-3691.

[79]张玉涛, 张梦凡, 史学强等. 人员安全帽佩戴轻量化检测方法研究[J]. 安全与环境学报, 2023, 23(02): 474-480.

致谢

行文至此，感慨万千。看似漫长的学生生涯即将结束了，三年的研究生生活即将结束，回首一路上的起伏，每一点都令人难忘。在此对关心我和支持我的老师和同学表示诚挚的感谢。

教诲如春风，师恩似海深。首先我要感谢我的导师王凤随教授，研究生生涯很幸运成为王老师的学生。老师渊博的专业知识，严肃的科学态度，严谨的治学精神，诲人不倦的高尚师德都对我产生了深远的影响。从课题的选择到最终完成，王老师都始终给予了我悉心的指导，不仅使我树立了远大的学术目标，还使我明白了许多为人处世的道理，同时还在精神和生活上给了我以无微不至的关怀。在此，谨向王老师老师致以我最诚挚的谢意和最衷心的感谢

恩同再造，债感无以为报。致谢我的爸爸妈妈及姐姐。感谢您们不计较回报的深情厚爱，始终尊重并支持我人生道路上的每一个选择，成为我走向未来的磐石般的后盾。有了您们，我有了无畏困难、勇往直前的勇气。在您们的光辉照耀下，我庆幸自己能不断前行。

笔走龙蛇，情深意重。对我的室友们以及课题组的所有同学，我亦怀有真挚的感谢与思念。感谢你们教会了我宝贵的生活哲学，是你们把乏味的学术生活变得绚丽多姿。特别感谢师兄师姐们在科研的道路上提供的无微不至的关照，无论是精心的指导、宝贵的建议，还是在写作论文时倾情相助，每一次的互助都使我们一同跨越了实验与学术的重重难关。

最后感谢百忙之中抽出宝贵时间参与论文评审和答辩的专家、老师们，向你们致以诚挚的谢意

攻读学位期间取得的学术成果情况

1 发表学术论文情况

[1] 杨海燕,王凤随,张兴旺.基于上下文增强和特征融合的目标检测算法[J/OL].重庆工商大学学报(自然科学版), 1-8

[2] 张兴旺,王凤随,杨海燕.基于多特征信息融合的疲劳驾驶检测算法[J/OL].重庆工商大学学报(自然科学版), 1-11

2 参与科研项目情况

（1）安徽省自然科学基金(2108085MF197)

（2）安徽高校省级自然科学研究重点项目(KJ2019A0162)

（3）安徽工程大学国家自然科学基金预研项目(Xjky2022040)