分类号：TP391　　学校代码：10363
密　级：公开　　学　　号：2220340102

安徽工程大学
Anhui Polytechnic University

硕士学位论文

题目　融合多尺度特征与信息增强的单阶段目标检测算法研究

论 文 作 者　邵凯丽
指 导 教 师　王凤随
学科（专业）　控制科学与工程
研 究 方 向　深度学习与目标检测

论文提交日期：2025 年 5 月 12 日

分类号：TP391　　单位代码：10363

密　级：公 开　　学　号:2220340102

题 目 融合多尺度特征与信息增强的单阶段目标检测算法研究

题 目 Research on Single-stage Target Detection Algorithm Based on Multi-scale Features and Information Enhancement

学生姓名：邵凯丽

指导教师：王凤随

专　　业：控制科学与工程

研究方向：深度学习与目标检测

论文答辩日期：2025 年 5 月 12 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：邵凯丽

日期： 2025 年 5 月 12 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：邵泓雨　　指导教师签名：王凤随

日期：　2025 年 5 月 12 日　　日期：　2025 年 5 月 12 日

融合多尺度特征与信息增强的单阶段目标检测算法研究

摘 要

随着深度学习技术的发展，利用卷积神经网络表征图像特征的方式被引入目标检测等领域，并在智能监护、公共安全等方面得到实际应用。然而，在实际应用场景中，目标检测模型受到光照、遮挡等环境因素的干扰，检测目标的特征信息提取不足，同时，深层网络中存在特征信息丢失的问题，两者都导致了检测效果不佳。此外，目标检测对象自身的形态变化也会对深度学习目标网络的检测效果产生影响。为了提高目标检测的准确性，论文选用单阶段目标检测算法 YOLOv7-tiny 作为基线检测器，对不同场景下的人体摔倒行为进行检测，通过多尺度特征融合和信息增强等方法，实现了特征信息的有效表达，并提高了高分辨率和强语义特征信息的交互能力，深、浅层特征信息得到了有效利用。其主要研究内容如下：

（1）针对轻量网络 YOLOv7-tiny 检测准确性不足的问题，提出了深度感知的多尺度目标检测算法。首先，通过引入空间感知检测头，增强网络捕获深层语义信息的能力，提升对多尺度姿态信息的处理效率，进而强化模型对目标的定位识别能力；然后，利用显示视觉中心模块，构建层内特征的全局依赖关系，聚焦局部特征信息，进一步提高网络的信息交互，实现检测准确性的提升。该算法在人体摔倒数据集上检测精度达到了 76.6%，较基线网络提高了 2.4%，实验结果表明

改进算法一定程度上提高了检测的准确性。

（2）针对 YOLOv7-tiny 算法在深层特征信息处理过程中检测效率降低的问题，提出了一种基于重参数化的自适应空间特征融合算法。首先，利用高效重参数化模块提升局部跨通道信息交互能力，增强浅层网络对人体姿势的局部信息的处理；然后，引入自适应空间特征融合结构，捕获人体姿态的全局信息，提高网络深层特征信息的融合能力；最后，优化回归损失函数，提高网络的置信度。该算法在人体摔倒数据集的平均精度达到了 77.2%，较基线网络提高了 3%，并且摔倒姿势的检测精度达到了 86.2%，实验结果表明改进模型有效提高姿势判别能力，并满足实时性的需求。

（3）针对深层网络的浅层特征信息的丢失问题，提出了联合信息增强和特征融合的改进 YOLOv7-tiny 算法。该方法通过在主干网络引入对比感知全局信息增强模块，增强对重要特征信息的敏感性，提取到更为丰富的语义信息，有效降低目标识别漏检的可能性。然后，为了充分利用不同层级的特征，引入融合 CA 注意力的密集坐标特征融合结构，提高不同层之间的信息交互，缓解信道信息聚合不充分的问题，提高网络对摔倒目标的检测精度。最终实验结果表明，该算法在人体摔倒检测数据集的平均检测精度达到了 77.0%，同时在学生课堂行为数据集 SCB1、SCB2 和 PASCAL VOC 2007 测试集的准确率分别达到了 78.4%、91.1%和 81.5%，具有泛化性。

关键词：卷积神经网络，目标检测，多尺度特征，特征融合，信息增

强

Research on Single-stage Target Detection Algorithm Based on Multi-scale Features and Information Enhancement

ABSTRACT

With the development of deep learning technology, the use of convolutional neural networks to characterize image features has been introduced into fields such as target detection and has been practically applied in intelligent guardianship and public safety. However, in real-world application scenarios, object detection models are often affected by environmental factors such as lighting conditions and occlusion, leading to insufficient extraction of the target's feature information. Additionally, there is a problem of feature information loss in deep networks, which results in poor detection performance. Moreover, variations in the shape or form of the objects being detected can also impact the effectiveness of deep learning-based object detection networks. In order to improve the accuracy of target detection, this paper chooses the image-based single-stage target detection algorithm YOLOv7-tiny as the baseline detector to detect the human fall behavior in various scenarios. By employing methods such as multi-scale feature fusion and information enhancement, it aims to enhance the representation of feature information, effectively utilizing both deep and shallow feature information. This approach improves the interaction between high-resolution and strong semantic feature information. The main research contents are as follows:

(1) To address the insufficient detection accuracy of the lightweight network YOLOv7-tiny, a depth-aware multi-scale object detection

algorithm is proposed. First, by introducing a spatial awareness detection head, the algorithm enhances the network's ability to process multi-scale pose information and strengthens its capacity to capture deep semantic information, thereby improving the model's target localization and recognition capabilities. Then, by utilizing an explicit visual center module, the algorithm constructs global dependencies within intra-layer features, focusing on local feature information to further enhance the network's information interaction and improve detection accuracy. The algorithm achieves a detection accuracy of 76.6% on the human fall dataset, which is a 2.4% improvement over the baseline network. Experimental results demonstrate that the improved algorithm enhances detection accuracy to a certain extent.

(2) Aiming at the problem that the detection efficiency is reduced when depth information is utilized, a re-parameterized improved YOLOv7-tiny algorithm is proposed. First, an efficient re-parameterization module is utilized to improve the network's detection efficiency, enhance local cross-channel information interaction, and strengthen the shallow network's ability to process local information related to human poses. Then, an adaptive spatial feature fusion structure is introduced to capture global information about human poses, improving the network's ability to fuse deep feature information. Finally, the regression loss function is optimized to enhance the network's confidence. The algorithm achieves an average accuracy of 77.2% in the human fall dataset, which is 3% higher than the baseline network, and the detection accuracy of the fall pose reaches 86.2%, and the experimental results show that the improved model effectively improves the pose discrimination ability and meets the real-time demand.

(3) Aiming at the problem of the loss of shallow feature information in the deep network, a dense coordinate feature fusion improvement

algorithm is proposed. By introducing the contrast aware global information enhancement module in the backbone network, it enhances the sensitivity of important feature information, extracts richer semantic information, and effectively reduces the possibility of target recognition leakage. Then, in order to make full use of the features of different layers, a dense coordinate attention feature fusion module is introduced to improve the information interaction between different layers, alleviate the problem of insufficient channel information aggregation, and improve the network's detection accuracy of falling targets. The final experimental results show that the algorithm achieves an average detection accuracy of 77.0% on the human fall detection dataset, as well as 78.4%, 91.1% and 81.5% on the student classroom behavior datasets 1 (SCB1), the student classroom behavior datasets 2 (SCB2) and the PASCAL VOC 2007 test set, respectively, which is generalizable and meets the real-time requirements.

KEY WORDS: convolutional neural network, target detection, multi-scale features, feature fusion, information enhancement

目录

第1章 绪论

1.1 研究背景与意义

深度学习作为人工智能领域的重要技术，从大量数据中自动学习复杂的特征表示，并在图像、语音、文本等任务中取得了突破性进展[1]。计算机视觉作为深度学习的核心应用领域之一，致力于让机器“看懂”图像或视频中的信息，其研究范围涵盖图像分类[2]、目标检测[3]、图像分割[4-5]、姿态估计[6-7]等多个方向。其中，目标检测是计算机视觉中的一项关键任务，其利用深度学习算法，识别出视频或图片中的特定对象，并确定其准确位置。目标检测的研究不仅具有重要的理论意义，还在安防监控[8-9]、医学诊断[10-11]、交通检测[12-14]、工业生产[15-18]等领域中广泛应用：

（1）安防监控。随着城市化进程的加快和公共安全需求的提升，智能监控系统替代了传统的人工翻阅的方式，利用计算机视觉对摄相机采集的视频图像进行智能处理，能够自动的分析、识别出异常行为[8]或可疑对象，大幅提高了安防效率。例如，在公共场所（如机场、车站、商场等）安装的智能监控系统，可以通过目标检测技术实时分析视频流，当识别出摔倒[9]等异常行为，系统立即发出警报，降低意外发生时人的身体损伤。

（2）医学诊断。医生依赖个人经验对 X 光片等医学影像诊断疾病，容易出现误诊或漏诊的情况，通过引入目标检测技术，计算机能够自动定位病灶区域，辅助医生做出更准确的诊断。例如，在癌症筛查中，通过对医学影像的分析，系统可以自动标记出可疑的病变区域，并提供详细的尺寸、形状和位置信息[10]。这种精确的定位帮助医生快速识别肿瘤区域，不仅提高了诊断的准确性，还大大缩短了诊断时间，为患者争取了宝贵的治疗时机。此外，在病理分析中，目标检测技术可以帮助医生快速识别组织切片中的异常细胞[11]，从而提高病理诊断的效率。同时，目标检测技术还在手术导航中发挥了重要作用，它能够实时定位手术器械和病灶区域，不仅提升了手术的精准度，也为远程手术提供了可能。

（3）交通检测。随着城市化进程的加快和城市车辆数量的增加，交通拥堵

和交通事故成为了亟待解决的问题。为应对这些问题，在路口安装智能摄像头实时监测道路状况，识别行人、车辆和交通标志，并统计车辆和行人的数量，并根据交通流量动态调整信号灯的时长，从而缓解交通拥堵。在智能交通系统中，这种实时数据分析不仅提高了道路使用效率，还减少了不必要的等待时间。此外，在自动驾驶领域，车辆通过摄像头、雷达等传感器感知周围环境，并利用目标检测技术实时识别出行人[12]、车辆[13]、交通标志[14]等关键对象，做出加速、减速、转向等驾驶决策，提高了行驶的安全性，还为未来的智能交通系统奠定了基础。

（4）工业生产。在工业生产领域，目标检测技术自动检测出不合格产品，提高生产效率和产品质量，为质量检测和自动化生产提供了强大的支持，被广泛应用在电子产品[15]、食品加工[16]、汽车制造[17]、纺织[18]等行业。在电子产品制造中，通过对电路板图像的分析，系统可以自动检测出焊接缺陷、元器件缺失等问题，并将不合格产品标记出来，大幅度缩短产品检测周期。在食品加工领域，目标检测技术可以用于检测食品的外观缺陷，如裂纹、变色等，从而确保食品的质量和安全。在汽车制造中，该技术可以用于检测零部件的尺寸和形状是否符合标准，确保生产标准一致性，从而提高整车的安全性和可靠性。在纺织行业中，目标检测技术可以用于检测织物的瑕疵，如污渍、破洞等，助力企业提高产品的市场竞争力。

随着深度学习技术的不断发展，目标检测的精度和效率得到了显著提升，但同时也面临着小目标检测、遮挡问题、实时性要求等挑战。因此，深入研究目标检测算法，不仅能够推动计算机视觉领域的理论创新，还能为实际应用提供更加高效、鲁棒的解决方案，具有重要的学术价值和社会意义。

1.2 国内外研究现状

目标检测可分为基于传统机器学习的方法和基于深度学习的方法。基于传统机器学习的方法检测过程如图 1-1 所示，图像在筛选候选框后，提取特征并完成判别和分类。传统的目标检测算法主要通过滑动窗口遍历图像，预估检测目标的存在区域，并使用基于边缘、纹理、颜色等手工设计特征提取图像特征。常用的图像特征有了哈尔特征[19](Haar-like Features, Haar)、方向梯度直方图

[20](Histogram of Oriented Gradient, HOG)、尺度特征不变性[21](Scale Invariant Feature Transform, SIFT)、局部二值模式[22](Local Binary Patterns, LBP)，其中，LBP 特征将图像的局部纹理特征转换为特征向量，Haar 特征用以反映预想的灰度变化情况，HOG 用以检测边缘特征，SIFT 适用于物体识别、影像追踪和动作比对等。

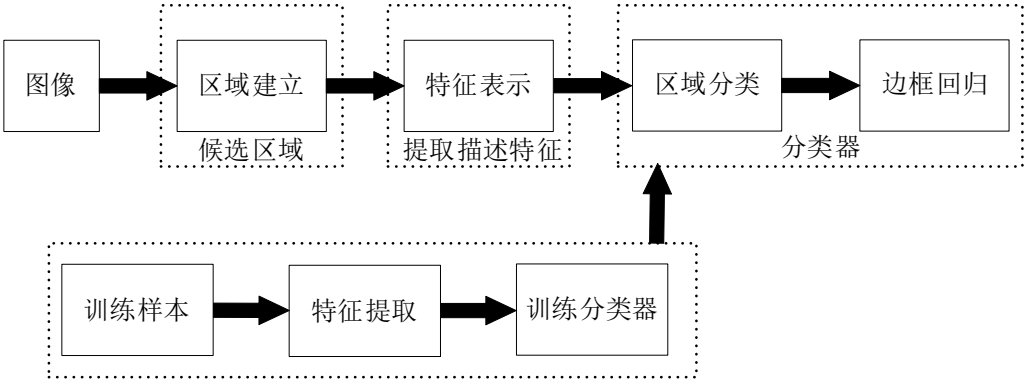


图 1-1 传统目标检测算法流程

在传统的目标检测算法中，通常使用针对特定类别进行人工设计的单一算子，这种方法在处理多目标检测任务时，模型的鲁棒性不佳，达不到很好的检测效果。同时，由于目标检测的尺度不一，传统目标检测方法在使用多尺度的滑窗进行图像遍历的过程中，产生了许多冗余窗口。相比之下，基于深度学习的方法通过深度学习的神经网络学习特征表示和对象分类算法，有效提高了目标识别算法的检测能力。随着技术的发展，根据检测任务不同的需求，目标检测方法逐渐多样化，涵盖了基于图像和基于点云的检测方式，并且每种检测方式可以应用于不同的场景。

1.2.1 基于图像的目标检测

基于图像的目标检测是最常见的目标检测方式，主要依赖于二维图像数据进行目标识别。根据传感器的不同，图像可以分为可见光图像和红外图像两种类型，其中，可见光图像保留了丰富的颜色和纹理信息，从可见光图像中捕捉目标的形状、颜色和纹理特征，可以实现高效的目标识别与定位，适用于大多数日常场景的目标检测任务。但在复杂环境下（如低光照、雨雾天气等），其应用受到限制。红外图像利用目标的热辐射特征，捕获物体发出的红外辐射并形成图像，因此可

以应用在夜间、强光等环境中。尽管如此，红外图像的分辨率较低，并且丢失目标部分颜色和纹理特征信息，带来了一定的局限性。

基于图像的目标检测算法分为单阶段目标检测算法和两阶段目标检测算法。其中，两阶段目标检测算法先生成候选目标区域，再对这些区域进行分类和检测位置调整；单阶段目标检测算法采用端到端的密集预测方式，直接在整个图像上生成目标的类别和位置信息，相较于两阶段目标检测算法，其速度上更具优势，适用于实时目标检测任务。

（1）基于图像的单阶段目标检测算法。采用滑动窗口或锚点的方式，在图像的不同尺度和不同位置上生成一系列的候选框，代表检测算法有 YOLO、SSD。其中，YOLO 算法将图像的边界框和物体类别的概率看作单一的回归问题，利用锚点和修正的方法获得检测结果，如表 1-1 所示。

表 1-1 YOLO 系列算法

模型	骨干网络(Backbone)	颈部(Neck)
YOLOv1[23]	Darknet-24	-
YOLOv2[24]	Darknet-19	-
YOLOv3[25]	Darknet-53	FPN
YOLOv4[26]	CSPDarknet-53	SPP+PANet
YOLOv5[27]	Focus+CSPDarknet	SPP+PANet
YOLOv6[28]	CSPDarknet	PANet
YOLOv7[29]	ELAN	SPPCSPC+PANet
YOLOv8	CSPDarknet-53	SPP+PANet
YOLOv9[30]	GELAN	SPP+PANet
YOLOv10[31]	CSPDarknet-53	SPP+PANet
YOLOv11	C3K2	SPP+PANet

SSD 算法[32]通过全卷积神经网络利用不同尺度的锚框进行多尺度目标预测，相较于 YOLO，其速度更快、精度更高，对小目标检测也更为友好，在此基础上，众多学者致力于增强 SSD 模型的语义信息，以提升检测效果。例如，DSSD[33]采用 Resnet101 替换 VGG16 作为主干特征提取网络，在预测阶段前引入残差模块，

并在 SSD 末端增加多个反卷积块，但其结构复杂，检测速率较低。FSSD[34]借鉴了 FPN 的思想，采用轻量级的特征融合模块，将不同层间、不同尺度的特征进行投影拼接，下采样模块获得新的特征金字塔，反馈给检测器得到最终的检测结果，但 FSSD 忽略了底层特征与高层特征之间的联系，使最后融合的特征不够充分，容易引入噪声和冗余信息。DF-SSD[35]利用 DenseNet 取代 VGG16，将残差模块和双向融合模块相结合，使得小物体具有更多的语义信息，但检测速度并不理想。

SSD 和 YOLO 基于锚框进行检测，而锚框的引入会导致正负样本不均衡，增加参数量。为了解决此问题，无锚点框的单阶段目标检测算法被提出。Law[36]等人提出了一种配对关键点的 CornerNet 算法，消除了单阶段算法对锚框的需要，整个网络的训练从头开始并不基于预训练的分类模型，设计新的池化方法更好地定位边界框的角点。为了改善 CornerNet 算法缺少物体外观特征信息，Zhou[37]等人提出了基于极值点的目标检测器 ExtremeNet，预测极值点热图和中心点热图。Tian[38]等人提出的 FCOS，无需对关键点检测，利用峰值点附近的图像特征回归得到目标的位置信息。2019 年，受 CornerNet 的启发，Zhou[39]等人提出了 CenterNet 模型，将目标检测问题转变为中心点估计问题，检测速度也更快。

（2）两阶段目标检测算法分为两个主要阶段，第一阶段是利用区域提议网络(Region Proposal Network, RPN)等方式生成候选目标区域；第二阶段则对这些候选区域进行分类与定位，代表算法包括 Faster RCNN[40]、Mask RCNN[41]等。2014 年，R-CNN(Regional CNN, R-CNN)[42]是首个将卷积神经网络(Convolutional Neural Networks, CNN)用于目标检测的模型，其核心是运用无监督的选择性搜索(Selective Search, SS)方法，该方法将输入图像具有相似颜色直方图特征的区域进行递归合并产生候选区域，然后从提取候选区域的特征图中提取特征，并利用支持向量机(SVM)分类器对这些特征进行分类，最终采用非极大抑制(Non-Maximum Suppression, NMS)处理分类结果。R-CNN 模型对输入图像缩放时，破坏图像长宽比，图像出现失，为了解决这一问题，空间金字塔池化网络(Spatial Pyramid Pooling, SPPNet)[43]被提出，其在 R-CNN 的基础上只进行一次全图的特征提取，然后从全图特征中进行截取每个候选区域对应的特征，并将这些特征送

入空间金字塔池化层，特征尺寸固定有效避免了因图像缩放带来的长宽比破坏问题。2015 年 Girshick 指出 R-CNN 和 SPPNet 特征提取和预测采用多阶段操作的方式限制了 CNN 性能的充分发挥，基于此提出了进阶版的 Fast R-CNN 算法[44]，利用感兴趣区域(Region of Interest, RoI)池化代替空间金字塔池化，并使用全连接网络代替之前的 SVM 分类器进行物体分类和检测框修正。针对 Fast R-CNN 利用 SS 提取候选区域耗费时间较长的问题，Ren 等人提出了 Faster R-CNN 算法，引入候选区域网络(Region Proposal Network, RPN)在输入图像中提取候选区域，在获取各个候选区域对应的特征图后，将特征图送入 Faster R-CNN 进行物体分类和位置回归。Mask R-CNN 算法基于 Faster R-CNN 的框架构成，在基础特征网络之后新增了全连接的分割网络，利用非线性方法确定非整数的位置，并添加 Mask 分支实现目标检测和分割任务。2017 年，旷视科技团队提出了一个基于 Resnet101[45]结构的 R-CNN 结构，利用全连接层(Fully Connected layer, FC)代替全局平均池化，并将 K×K 的卷积拆分成 1×K 和 K×1 的卷积减少模型的参数量，大大减少了模型的参数量。

1.2.2 基于点云的目标检测

基于点云的目标检测利用几何分析和数学统计方法，处理传感器采集到的空间信息，生成目标的场景语义信息和空间信息，在自动驾驶[46]等领域具有重要意义。传统的点云检测方式基于点云的曲率、密度等属性提取特征，依赖聚类方法，鲁棒性不高，逐渐被深度学习算法所替代[47]。深度学习方法中，研究人员常用的点云数据的处理方法可以分为直接点云处理方法、基于体素的方法和基于点云投影的方法。

(1)直接点云处理方法。2017 年，Qi[48]等人使用最大池化等方式解决点云无序性问题，提出了点云直接处理的方法，实现直接从 3D 点云中提取特征信息的目标检测。为了提高局部特征的捕捉能力，Qi[49]等人在 PointNet 的基础上提出了 PointNet++，引入分层特征学习机制，以提取不同尺度下的特征，且通过在输入点云中选择中心点并围绕其划分局部区域，对局部区域进行采样和分组捕捉局部几何信息，提高了对点云局部特征的建模能力。

(2)基于体素的方法。为了提高点云数据处理速率，VoxelNet 算法[50]中提出了体素的概念，点云数据转换为大小规则的体素网格(3D voxel grid)，每个体素内的点会被处理成特征向量，再使用三维卷积神经网络进行处理，通过将密集的点云数据转换为体积表示的形式，简化数据表达。点云稀疏时，体素网格中存在较多空的体积单元，Yan[51]等人在 SECOND 中引入稀疏卷积处理非空体素，联合正弦值回归和方向分类训练，优化角度损失。2019 年，Lang[52]等人提出了基于点和体素的 PointPillars，点转化成柱体构成伪图片数据后，应用 2D 卷积层，实现端到端学习，点云检测速率进一步提高。

(3)基于点云投影的方法。Frustum-PointNets[53]利用 2D 目标检测器在图像中确定带有目标的 2D 兴趣区域，然后将该区域投影到 3D 空间中，形成点云中的棱锥，之后在棱锥区域进行目标检测以减少点云的处理量。AVOD[54]将点云的不同投影视角与图像结合，从不同视角和不同数据源提取特征，以提高检测性能，但其使用不同数据源特征融合时会导致数据处理更加复杂，增加了计算量。为了解决不同传感器融合特征不匹配的问题，BEVDetNet[55]将点云投影到 2D 鸟瞰图(Bird’s Eye View，BEV)中，只在 BEV 中进行目标中心回归以及朝向 3D 框的预测。除了使用点云鸟瞰图投影进行目标检测之外，FVNet[56]将点云投影到圆柱面上以生成保留丰富信息的距离视图(Range View，RV)，从 RV 中生成 3D 感兴趣区域，之后从该区域的点云中检测出感兴趣目标。基于点云投影得到多视图的方法简化数据表达方式，降低计算量，但 3D 点云投影到 2D 视图中会丢失部分信息，尤其是在单个视图中进行目标检测时会导致漏检错检率升高，而结合多视图的特征融合方法因为数据可能存在不同分辨率、视角和形式所导致对齐和配准的问题，从而影响特征融合效果，处理过程也更为复杂。

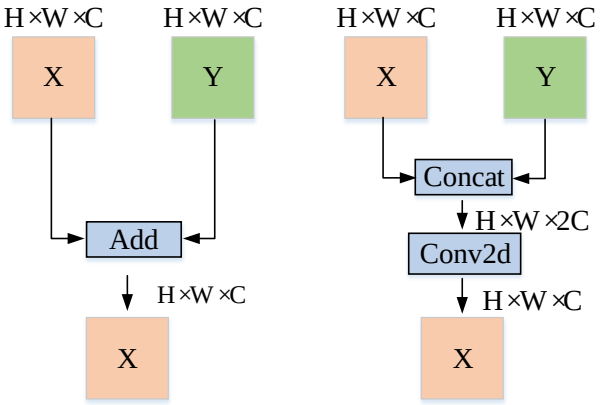
表 1-2 基于点云的单阶段目标检测算法

算法	创新处
SA-SSD[57]	辅助网络将卷积特征转为点级表示后执行前景分割和中心点估计； Part-sensitive Warping(PSWarp)解决检测框与置信度不匹配问题。
3DSSD[58]	特征采样策略保留内部点； 基于候选生成层、无锚点回归头和三维中心性分配策略的预测网络提高保

	留点的利用。
SE-SSD[59]	形状感知数据增强(Shape-Aware data augmentation, SADA)生成真实目标；方向感知距离损失(Orientation-aware Distance-IoU, OADI)优化。
HVPR[60]	自适应多尺度特征金字塔(Attentive multi-scale feature module, AMFM)。
PillarNeXt[61]	扩张卷积扩大感受野，引入空洞空间金字塔池化。
CoCa3D[62]	协作深度估计(Collaborative depth estimation, Co-Depth)消除深度歧义；协同检测特征学习(collaborative detection feature learning, Co-FL)共享高置信度检测特征。

1.2.3 特征融合及信息增强技术

卷积神经网络具有局部感知、权重共享、平移不变性等特性，能够有效地从输入数据中提取出有意义的特征，但是，在实际应用中，目标检测任务面临光照变化、遮挡、姿态变化等挑战，仅依靠单一尺度或层次的特征往往不足以应对复杂的目标检测任务。为了进一步提升模型的性能，特征融合和信息增强技术应运而生，通过结合不同层次、不同尺度的特征，以及引入额外的信息源，能够显著提高目标检测网络的表达能力和鲁棒性。



(a) 求和　　(b) 拼接

图 1-2 特征结合方法

在目标检测任务中，对输入图片逐层卷积，然后使用最后一层特征图进行预测，如 SPP-Net、Fast R-CNN 等就是采用这种方式。物体的大小和形状可能差异很大，小目标需要更为精细的低级特征（如边缘、纹理），而大目标则依赖于更

高层次的语义特征（如物体形状、类别），单个尺度的特征往往无法全面描述目标。因此，特征融合的目标之一是将不同尺度的特征结合起来，以捕捉到目标的全局和局部信息，图 1-2 中为求和及拼接两种方法结合特征的方式。

基于特征融合的方法提高了特征图利用率，ResNet 残差网络及其后续改进网络中，身份映射特征和残差学习特征通过短跳跃连接融合为输入，从而可以训练非常深的网络，但是，深度网络中多尺度特征信息的有效处理仍未解决。目标检测网络中的浅层卷积层主要捕获低级特征，而深层卷积层则捕获更高级的语义特征，通过将浅层和深层特征进行融合，模型可以在不同尺度上同时捕捉到目标的细节和全局信息，从而提高检测的准确性。Lin[63]等人基于 Faster R-CNN 提出了特征金字塔网络(Feature Pyramid Network, FPN)，通过自顶向下的路径将高层特征传递到低层，增强了低层特征的语义信息，使得模型能够更好地处理多尺度目标。FPN 的核心思想是通过横向连接(lateral connections)将高层特征与低层特征进行融合，保持低层特征空间分辨率，同时提升其语义表达能力。Liu[64]等人在 PANet 中提出一种基于 FPN 的自下而上的路径增强特征层次，以增加深层的低级特征信息，并提出自适应特征池化，通过聚合所有级别的特征以进行更好的预测，根据目标的大小和位置动态调整特征池化的区域，提高检测的准确性。NAS-FPN[65]通过神经架构搜索技术自动设计特征金字塔网络的方法，利用强化学习或进化算法等自动化搜索策略，从大量可能的网络结构中寻找最优的特征融合方式，从而提升多尺度目标检测的性能。Tan[66]等人提出了双向特征金字塔网络(Bidirectional feature pyramid network, BiFPN)，在 FPN 的基础上进一步优化了特征融合的方式。BiFPN 通过双向（自顶向下和自下而上）的特征传递路径，这种双向结构不仅增强了特征的表达能力，提高了模型在多尺度目标检测任务中的性能，还引入了加权特征融合机制，通过学习不同特征层的重要性权重，进一步优化了特征融合的效果。上述方式通过实现特征融合路径，从而提高低级特征信息和高级特征信息的有效利用，进一步实现检测能力的提高，但是增加了冗余信息。TridentNet[67]删除了特征金字塔的结构，并创建了具有不同感受域的多个尺度特定分支，以采用尺度感知训练和推理。

目标检测中，分类任务需要判断出检测目标的类别，需要高级的语义信息，

而定位任务需要确定目标的位置，更需要的是低级的空间特征。特征融合将不同尺度的特征结合起来，模型可以更好地理解目标的上下文信息，实现更好的协同效应。在一些复杂场景中，通过融合不同来源的特征，不仅可以提高模型的表达能力，还可以增强其鲁棒性。例如，红外图像和可见光图像[68]的融合可以弥补单一模态的不足，提升模型在低光照条件下的检测性能；点云数据和图像数据[69-70]的融合可以结合两者的优点，提升模型在复杂环境中的鲁棒性。

深度学习中，信息增强通过各种技术手段提升模型对输入数据的理解能力，从而改善模型的性能，可以分为数据层面、特征层面和模型层面。

（1）数据层面。深度神经网络通常需要大量的训练数据以避免过拟合，而标记数据非常有限。数据层面信息增强[71]通过对输入图像进行处理，解决着训练数据不足、训练数据类别不均衡的问题，其本质是通过增加数据的多样性或丰富性，提升模型的泛化能力。针对图像数据，可以利用旋转、缩放等操作，模拟不同的拍摄角度和距离，通过亮度调整、对比度调整等方式，模拟不同的光照条件，使用 CutMix[72]（图片中的一部分替换为另一张图片的对应内容）、Cutout[73]（随机裁剪）、Mixup[74]（混合标签）等混合样本的方法，通过将多个样本的图像和标签进行混合，生成新的训练样本，帮助模型学习到更多样化的特征。此外，训练 GAN[75]等生成对抗网络，合成新数据也是数据增强方式之一。

（2）特征层面。特征信息的有效利用对于深度学习模型的准确性至关重要，通过改进特征提取或特征表示可有效提高模型的表现。特征信息增强方式包含注意力机制和特征增强模块。注意力机制[76]允许模型在不同的特征图之间分配不同的权重，从而在处理信息时更加关注突出重要的区域，忽略不重要的区域，进一步提高检测准确率和效率。2014 年，Bahdanau[77]等人首次提出在 RCNN 上应用注意力机制进行图像分类。2018 年，Hu[78]等人提出了通道注意力，通过学习通道间关系提高特征的表达能力，Woo[79]等人在 CBAM 中提出先通道后空间的架构，进一步提高检测能力。2020 年，Carion[80]等人在 DETR 的编码器部分应用自注意力机制[81]对输入图像进行特征提取，解码器中将目标与图像特征对齐。特征增强模块包含残差连接、稠密连接等。残差连接(Residual Connection)通过将前一层的特征直接传递到后一层，而如 DenseNet[82]等稠密连接则是将每一层的特

征传递给后续所有层，这种增强方式通过提高特征的复用性促进信息的跨层流动，解决深层网络中的梯度消失问题，提升了模型的训练效果。

（3）模型层面。模型增强技术利用多任务学习、知识蒸馏和模型集成等方法通过改进模型结构或训练策略来提升性能。多任务学习通过引入多模态信息、时序信息、上下文信息等外部信息，提升模型的性能。例如，Mask R-CNN 结合目标检测和实例分割的任务框架，通过引入分支网络分别处理检测和分割任务，并将两者的结果进行融合，不仅可以精确定位目标，还可以生成高质量的分割掩码，适用于复杂的场景；在基于视频的目标检测任务中，模型可以利用时间序列信息跟踪目标并预测位置，有效提升检测的准确性和网络的鲁棒性。

综上所述，特征融合和信息增强技术在目标检测任务中扮演着至关重要的角色，通过将不同尺度、不同任务以及不同来源的特征信息进行处理，模型能够更全面地理解目标信息，从而在复杂场景下实现更精确、更鲁棒的检测性能。

1.3 论文的结构安排

基于研究的主要内容，论文一共包含六个章节，章节内容概述如下：

第一章为绪论，首先阐述了目标检测研究的时代背景和研究意义，其次概述了国内外目标检测的研究现状，总结了特征融合及信息增强技术，最后言简意赅的梳理了论文的主要研究内容。

第二章对论文研究涉及到的单阶段目标检测网络的相关技术进行介绍，具体包括 CNN 的基础理论、特征处理及分类回归，概述 Yolov7 算法和 Complex-YOLOv11 算法，为后文的研究提供基础理论和技术支持。

第三章、第四章、第五章从不同角度，利用特征融合和信息增强方式，提高单阶段目标检测算法的检测性能。

第六章对论文工作进行总结，根据研究内容展望未来的工作。

第 2 章 单阶段目标检测网络的相关技术分析

2.1 引言

与传统方法相比，深度学习方法利用神经网络学习特征表示，有效提高了目标检测技术的检测能力。本章主要概述目标检测的基础理论，对特征处理和目标分类与回归涉及的理论进行介绍，并对 2D 目标检测、3D 目标检测涉及的相关重点理论以及目标检测的数据集和评价指标进行阐述，为后续几章研究工作提供基础理论指导。

2.2 目标检测的基础理论

CNN 是一种前馈神经网络，主要包括输入层(Input)、卷积层(Convolutional Layer)、池化层(Pooling)、激活层(Activation Function)、全连接层(Fully Connected Layer)以及输出层(Output)。卷积神经网络是针对图像设计的，主要特点在于卷积层的特征是由前一层的局部特征通过卷积共享的权重得到。在卷积神经网络中，输入图像通过卷积和池化等一系列操作进行特征提取，然后低级特征逐步转换为高级特征。高级特征在经过全连接层和输出层进行特征分类，得到特征向量，表示当前输入图像的类别，如图 2-1 所示。

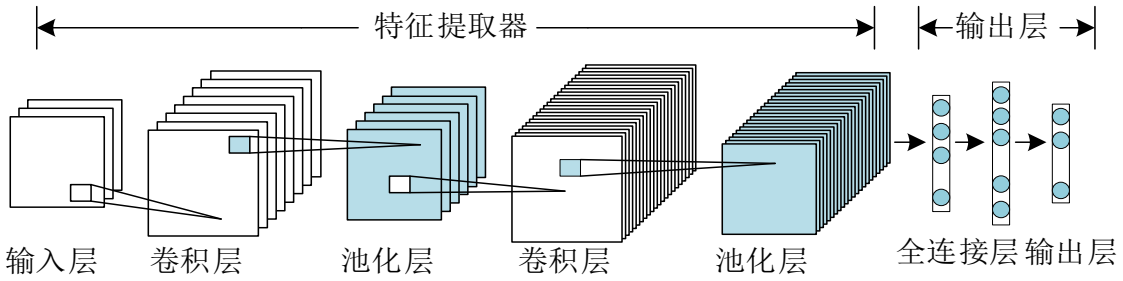


图 2-1 CNN 基础结构

(1)卷积层。以指定大小的卷积核在输入的特征图上以固定步长滑动，分别在多通道上进行局部连接，并通过多次迭代得到由低级到高级的不同层次深度的特征，相关权重可以通过反向传播算法优化得到。

(2)激活函数。对卷积输出进行非线性求解，增强函数映射能力，如 Sigmoid

函数、Tanh 函数、Relu 函数、Gelu 函数等。

(3)池化层。在特征通道里对特征图进行下采样，减小尺度大小。常用的池化有最大池化、均值池化、高斯池化等。

(4)全连接层。全连接层中的每个神经元与所有像素全连接操作，将局部特征整合成全局特征，用以计算每一类别的得分。

输入图像经过骨干网络进行多层次的卷积操作，捕捉图像中的局部特征（如边缘、纹理），提取高级语义信息（如物体形状、对象）。在这个过程中，目标检测网络生成了多个不同尺度的特征图，网络的检测头以特征图作为输入，并执行分类和回归两个主要任务。

2.2.1 特征处理

在目标检测任务中，有效的特征提取和处理是模型能够从输入数据中捕捉到有助于识别和定位目标的关键信息的基础。因此，特征信息的质量直接影响模型的性能。传统的特征提取方法，如 SIFT 和 HOG，依赖于手工设计的特征描述符。然而，随着深度学习的快速发展，CNN 通过自动学习高层次抽象特征，显著提升了特征信息的处理效果。

目标检测通常利用卷积及池化等方式处理特征，其中，卷积操作是深度学习中用于特征提取的核心工具，尤其在 CNN 中发挥了重要作用。在数学上，卷积(Convolution)是通过两个函数 f 和 g 生成第三个函数的一种数学运算，即一个函数在另一个函数上的加权叠加，其本质是一种特殊的积分变换。具体来说，对于两个实变函数 $f(\mathrm{t})$ 和 $g(\mathrm{t})$，它们的卷积 $(f*g)(t)$ 定义如下：

$$(f*g)(t)=\int_{-\infty}^{+\infty} f(\tau)g(t-\tau)d\tau \tag{2.1}$$

其中，$f(\tau)$ 表示被积函数，$g(t-\tau)$ 表示卷积核函数。

在图像处理中，图像可以理解为一个离散函数，而卷积操作可以理解为将卷积核在图像上滑动，卷积核沿水平或垂直方向移动到图像的不同位置。在每个位置上，卷积核与图像上对应区域进行逐元素相乘，所有的乘积结果相加得到了该位置的输出值。通过卷积计算，生成了新的图像，即特征图。以二维数组为例介绍卷积操作，如图 2-2 所示。

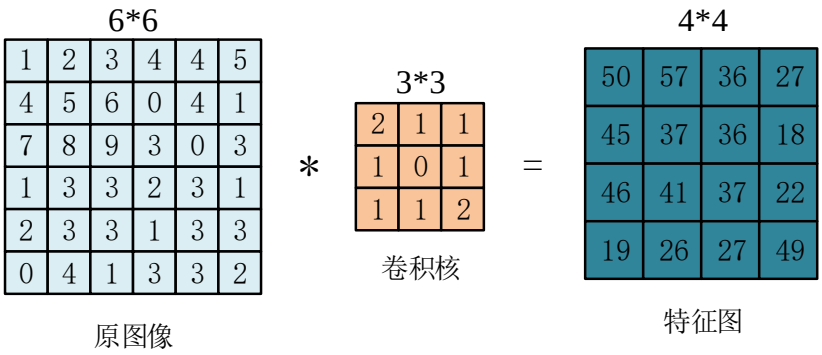


图 2-2 二维卷积计算过程

由图 2-2 所示，输入图像为一张 6×6 像素的黑白图片，通过与一个 3×3 大小的卷积核进行卷积运算，得到了一个新的 4×4 大小的黑白图片。卷积核是一个小矩阵，在与输入图像的局部区域进行逐元素相乘并求和的过程中，不同的卷积核权重大小关系着所提取的特征类型，不同的权重组合可以提取边缘、纹理、形状等不同类型的特征。卷积具有线性、平移不变性和参数共享等性质，卷积网络能够从输入数据中捕捉到丰富的局部和全局信息，逐步抽象出更有意义的特征。

池化是卷积神经网络中的重要操作之一，主要用于对输入特征图进行下采样，以减少数据的空间尺寸（宽度和高度），从而降低计算复杂度，同时增强模型对平移、旋转等变换的鲁棒性。池化窗口在输入特征图上滑动，对于每一个局部区域计算相应数值，池化操作与卷积层交替使用，逐步提取高层次的特征，减少数据的空间尺寸。常见的池化操作包括最大池化和平均池化。最大池化选择局部区域中的最大值作为输出，能够保留最显著的特征，而平均池化计算局部区域中的平均值作为输出，能够平滑特征图，如图 2-3 所示，滑动窗口的尺寸大小为 2×2，滑动的步长为 2，通过最大池化和平均池化，特征图尺寸大小减小。

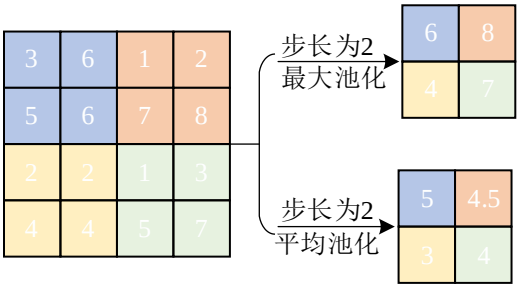


图 2-3 两种池化的操作过程

卷积和池化操作都可以改变特征图的尺寸大小，如表 2-1 所示，其中 F_1 表示输入的特征图大小，F_2 表示处理后的特征图大小，滑动窗口大小为 k，卷积的零填充量为 P，步长为 s，步长即滤波器每次移动时移动的偏移数量。

表 2-1 特征图尺寸

处理方式	特征图
卷积	$F_2=\frac{F_1-k+2*P}{s}+1$
池化	$F_2=\frac{F_1-k}{s}+1$

2.2.2 分类与回归

目标检测网络的核心任务包括分类和回归。两阶段目标检测算法通过合并相似的像素块的选择搜索(Selective Search)方法或者利用卷积神经网络直接预测候选区域的区域候选网络(Region Proposal NetWork, RPN)方法生成候选区域，单阶段目标检测算法则省略了这一步骤，直接对整个图像进行目标检测。无论是单阶段还是两阶段目标检测算法，在得到候选区域后都需要执行两个基本任务：分类和回归。分类任务用于确定目标所属的类别，而回归任务则是用于精确预测目标的位置，通常以边界框(Bounding Box)表示。

目标检测网络的检测头接收由主干网络处理后的特征图作为输入，利用全连接层或多层感知机(Multilayer Perceptron, MLP)对每个候选区域进行分类，在单阶段检测器中可能是直接从特征图上定义的锚点框中进行分类。特征图上的每一个点都包含了丰富的上下文信息，使得模型可以判断该点对应的区域内是否存在感兴趣的目标以及具体是什么类型的对象，实现检测分类的准确性。同时，检测头也会对边界框进行调整，实现更为精确地定位目标。计算预测边界框与真实边界框之间的差异，并通过训练最小化这种差异，具体来说，模型会输出相对于预设锚点框的偏移量，其包括中心坐标（x，y），宽度 w，高度 h 的变化，然后使用这些偏移量来微调初始的锚点框位置，使之更加贴合实际目标，实现定位的准确性。偏移量的计算过程如下：

(1)假设检测的锚点框为(x_a, y_a, w_a, h_a)，而真实边界框为(x^*, y^*, w^*, h^*)。

(2)计算回归器预测的偏移量，如以下公式所示：

$$\Delta x=\frac{x^*-x_a}{w_a} \tag{2.2}$$

$$\Delta y = \frac{y^* - y_a}{h_a} \tag{2.3}$$

$$\Delta w = \log\left(\frac{w^*}{w_a}\right) \tag{2.4}$$

$$\Delta h = \log\left(\frac{h^*}{h_a}\right) \tag{2.5}$$

(3)通过偏移量调整锚点框的位置和大小，得到新的锚点框数据，如以下公式所示：

$$x = x_a + w_a \cdot \Delta x \tag{2.6}$$

$$y = y_a + h_a \cdot \Delta y \tag{2.7}$$

$$w = w_a \cdot e^{\Delta w} \tag{2.8}$$

$$h = h_a \cdot e^{\Delta h} \tag{2.9}$$

模型训练过程中，通常定义一个合适的损失函数，衡量预测结果与实际标签之间的差距。常见的损失函数包括交叉熵损失（用于分类任务）和 Smooth L1 Loss 或 IoU Loss（用于回归任务），通过对损失函数的优化，可以逐步调整模型参数，提高检测精度。在训练的过程中可能会出现了一个目标产生多个重叠的检测框的情况，因此需要一种机制来去除冗余的检测结果。非极大值抑制(Non-Maximum Suppression, NMS)是一种常用的技术，它基于预测得分对检测框进行排序，并删除那些与得分最高框重叠超过一定阈值的其他框。

2.3 目标检测算法

2.3.1 YOLOv7 算法

YOLOv7 是 Alexey Bochkovskiy 团队提出的 YOLO(You Only Look Once)的算法，其在结构上进行创新，设计了高效远程聚合网络(Extended Efficient Layer Aggregation Networks, ELAN)进行特征提取，以轻量化的设计进行高效的特征提取和多尺度预测。其结构如图 2-4 所示。

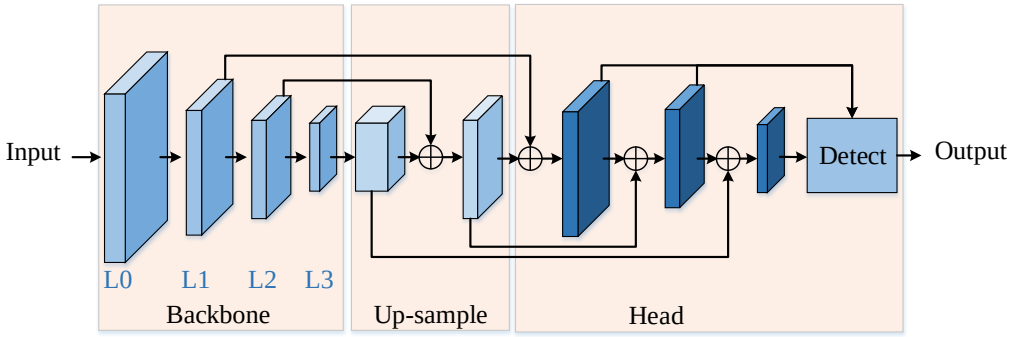


图 2-4 YOLOv7 的结构

YOLOv7 中利用池化层和上采样模块得到多尺度特征图，并使用 ELAN 提取并处理多尺度特征。ELAN 中使用了多个连续的深度卷积层，分割基础层的特征映射并将其重组为新的模块，通过最短最长的梯度路径提高网络的学习能力，促进信息流的传递并缓解深层网络中的梯度消失问题。为了满足边缘设备的需求，提出轻量的 YOLOv7-tiny，通过减少 ELAN 的卷积层数，降低多尺度特征变化的计算，有效提高边缘计算的效率。轻量级 ELAN 的结构如图 2-5 所示，其中 CB 为卷积层，Concat 为按通道特征拼接操作。

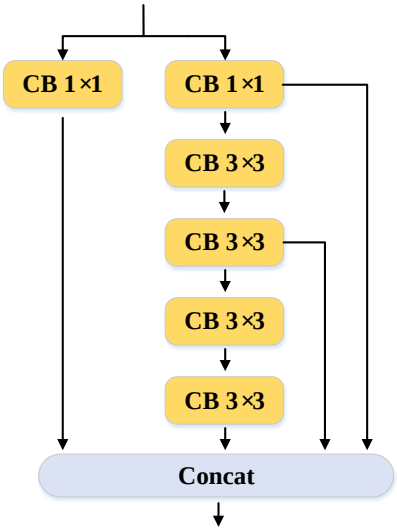


图 2-5 轻量级 ELAN 的结构

2.3.2 Complex-YOLOv11

Complex-YOLO 中为点云信息处理提供了新范式，结合 YOLOv11 优化其检测能力，其主要流程如图 2-6 所示。

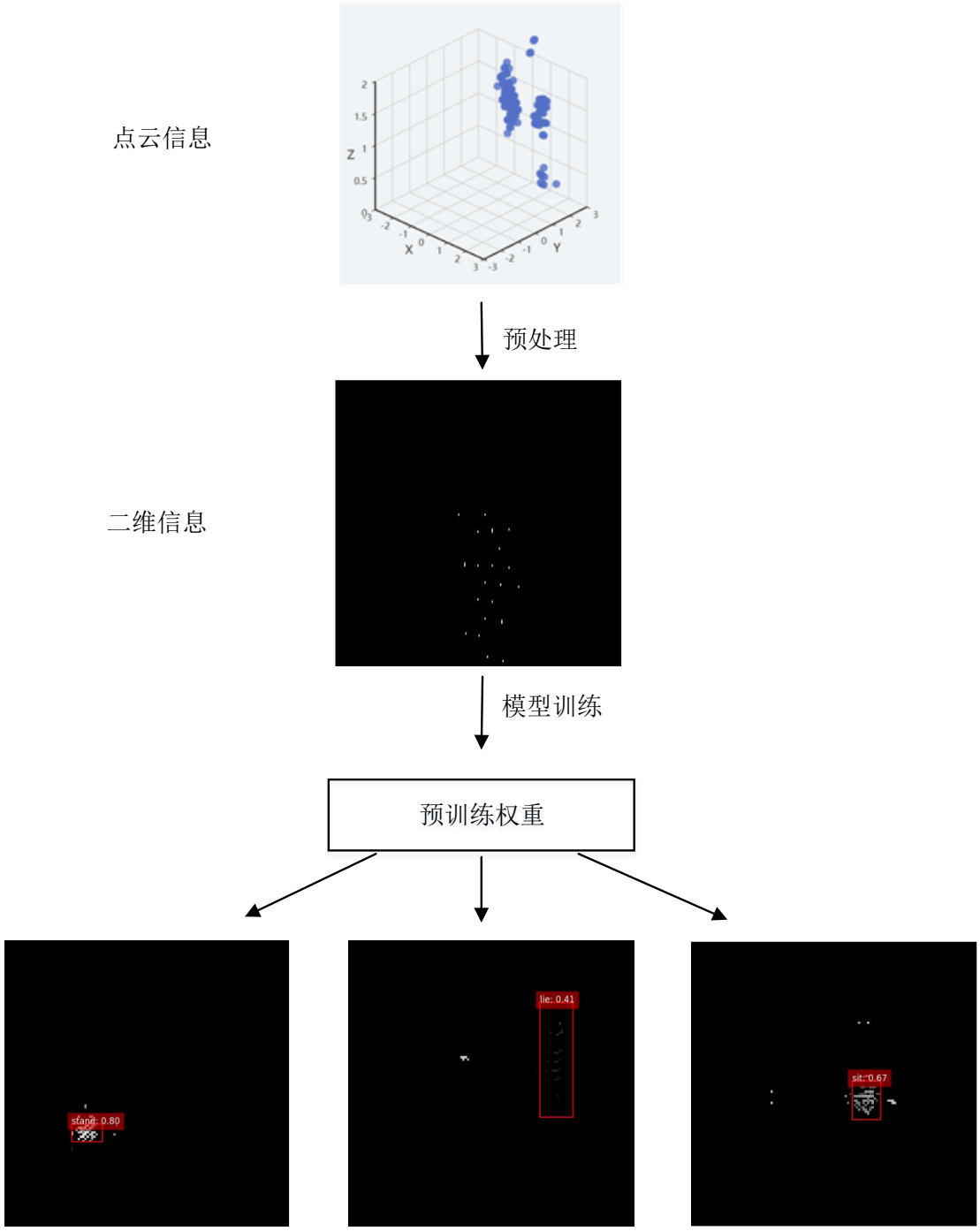


图 2-6 Complex-YOLOv11 算法流程

由图 2-6 所示，先将保存的激光雷达点云信息转变为二维特征信息。三维信息转变二维信息时，为了保留其深度信息，其转换为强度信息。然后利用 Yolov11 对处理好的数据集预训练，学习多尺度特征信息，得到预训练权重。最后将 Yolov11 的检测权重应用到实时推理任务中，可分析其检测结果。

2.4 数据集和评价指标

选择合适的数据集和评价指标对于目标检测算法的研究至关重要。通过使用标准化的数据集和科学的评价体系，不仅可以促进学术交流和技术进步，还能帮助开发者更好地理解和优化自己的模型。

2.4.1 目标检测数据集

目标检测方法有很多，为了评估和比较不同的目标检测算法，研究者开发了一系列标准的数据集。Pascal VOC 数据集和 COCO(Common Objects in Context)数据集是两个广泛使用的基准数据集：Pascal VOC 是一个经典的图像目标检测数据集，自 2005 年起每年举办一次挑战赛。VOC 数据集涵盖了 20 个常见类别，每个类别包含一定数量的训练、验证和测试图像。尽管 VOC2012 之后没有再发布新的测试集标签，但它仍然是衡量算法性能的重要参考之一。COCO 数据集相比 Pascal VOC 数据集，COCO 拥有更多的类别（80 类）和更大的图像规模（约 12 万张训练/验证图像）。COCO 不仅支持目标检测任务，还扩展到了实例分割、关键点检测等多个领域，成为当前最流行的目标检测数据集之一。

除了上述通用数据集外，针对特定应用领域的需求，还可以构建定制化的数据集。例如，针对密集的小尺度目标检测研究时，数据集应当注意检测目标是否符合小尺度以及密集的要求。本论文针对复杂环境中摔倒姿势的检测问题进行研究，所使用人体摔倒姿势识别数据集图片来源于 2019 年中国华录杯数据湖算法大赛定向算法（人体摔倒姿势识别）所提供的图片资源。为了模拟公开场所的检测环境，数据集包含了低光照、遮挡等情况，如图 2-7 所示。

(a)目标遮挡

(b)光照不足

(c)干扰目标多

(d)检测目标小

图 2-7 摔倒检测场景示例

图像中的检测对象分为摔倒姿势和其他姿势（如蹲下、站立等姿势），以便训练和测试的摔倒检测模型。另外，在网络模型训练数据集时，应当注意确保样本的多样性和代表性，遵循合理的划分原则，本论文在实验过程中将数据集的 7783 张图像按照 8：1：1 随机划分为训练集、验证集和测试集，其中，测试集的 778 张图片中包含 2876 个检测对象，897 个摔倒的检测对象，1979 个其他姿势的检测对象。

表 2-2 数据集划分

数据集/张	训练集/张	验证集/张	测试集/张
7783	6226	779	778

2.4.2 目标检测评价指标

在目标检测任务中，常用的评价指标包括以下几个方面：

（1）精确率(Precision)和召回率(Recall)：这两个指标分别衡量了模型预测结果的质量和完整性。精确率表示所有被预测为正类的目标中有多少是真正的正类；召回率则反映了所有实际为正类的目标中有多少被正确预测。两者之间的平衡可以通过 F1 分数来评估。

（2）精度(Average Precision, AP)和平均精度(mean Average Precision, mAP)：AP 是指某一类别下 Precision-Recall 曲线下的面积，而 mAP 则是所有类别 AP 值的平均值。这是目前最常用的目标检测评价指标之一，尤其在 COCO 等大型数据集上得到了广泛应用。

（3）交并比(Intersection over Union, IoU)：IoU 用来衡量预测框与真实框之

间的重叠程度，通常作为判断两个边界框相似性的依据。在计算 AP 时，IoU 阈值的选择非常重要，不同的数据集可能采用不同的标准。

2.5 本章小结

本章简要对深度学习技术所涉及的基础理论进行阐述，首先对深度学习卷积神经网络所关联的基本概念进行了分析说明，并介绍了目标检测任务中特征处理、目标分类与回归的基础内容。此外，介绍了 2D 检测任务中的 YOLOv7 算法，并结合 Complex-YOLO 介绍了融合 3D 点云信息的目标检测算法，阐述了目标检测算法的数据集与评价指标。

第 3 章 深度感知的多尺度目标检测算法

3.1 引言

人在意外摔倒后身体受到影响，引发一系列健康问题，严重时将威胁生命。据世界卫生组织统计报告，每年由摔倒致使人员死亡达到了 64.6 万人，受伤严重的超过 3730 万人次[83]。针对摔倒这一重要的公共健康问题，许多研究人员提出自动检测跌倒的方案[84-85]，迅速发现摔倒情况，伤者得到及时救助，从而减少摔倒带来的身体伤害。其中，人的摔倒检测[86-87]可以分为基于传感器的检测方法和基于计算机视觉的检测方法。基于传感器的方法通过随身携带的设备，得到被检测对象的运动信息[88]，或者分析人体运动引起的物理信号[89-90]变化，实现摔倒行为的检测目的，此检测方式需佩戴仪器，易受到周围环境的干扰，不利于突发的摔倒行为检测，而基于计算机视觉[91]的检测方法通过从检测环境中获取动态的行为信息，更适用于突发情况的摔倒检测。本章使用 YOLOv7-tiny 算法进行摔倒检测，轻量化的模型架构在目标检测任务过程中可以满足实现检测的需求，但是检测的准确性有待提高。为了优化这一问题，本章提出了一种深度感知的多尺度目标检测算法，对多尺度特征图进行目标检测，通过引入感知检测头，将高级语义信息深度提取，提高检测的准确性。

3.2 算法原理

YOLOv7-tiny 使用池化的方式降低特征图维度后，通过轻量化的 ELAN 模块获取特征信息，检测的速度大大提高。YOLOv7-tiny 的检测头将深度信息进行处理，预测特征信息，但是在面对复杂场景下人行为姿势缺失等情况下，Head 部分将 Upsample 提取的特征图在通道维度拼接，有效利用了浅层图像的特征信息，忽略了深层的强语义特征信息，目标检测能力和特征学习能力有所降低，降低了检测精度。本章使用空间感知检测头(Spatial Awareness, SA)和显示视觉中心模块(Explicit Visual Center, EVC)。如图 3-1 所示，改进算法先利用深度感知模块提高

深度特征图的表达能力，提高空间感知检测头分类与回归的准确性，然后，引入 EVC 模块提高多尺度信息的交互能力，增强小尺寸特征图的表达。

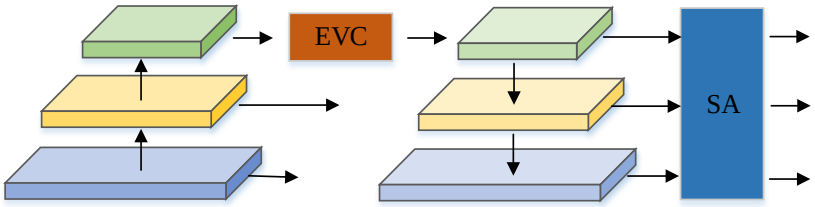


图 3-1 改进后的算法

3.2.1 空间感知检测头

检测头为目标检测网络提供检测目标的位置、大小和类别，是否有效处理特征网络提取的特征图影响着检测的准确性。注意力机制关注重要特征信息，进而提高信息的处理能力，本章节通过引入深度感知检测头，利用深度感知模块(Deep Perception Module, DPM)增强回归特征图的表达能力，实现检测准确性，其结构如图 3-2 所示。

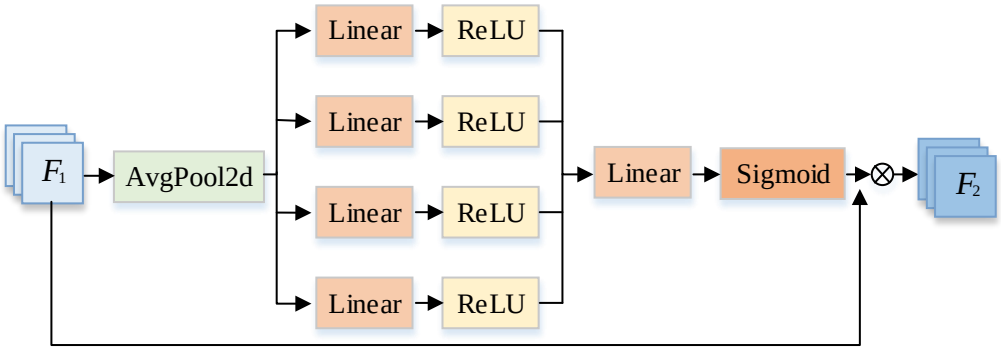


图 3-2 DPM 模块

如图 3-2 所示，检测头对不同感受野的特征信息进行处理和凝聚获取输入张量尺寸，获取输入特征图的批量大小 b 和通道数 c。使用全局平均池化将输入特征图压缩为一个 1x1 的特征向量，并重塑为(b，c)的形状，再通过四个全连接层处理特征向量，每个全连接层将特征向量维度变化后，通过 ReLU 激活函数进行非线性变换。生成四个中间特征向量，并将四个中间特征向量在通道维度上拼接，形成一个新的特征向量 $\boldsymbol{Y}$，其计算过程如下：

$$\boldsymbol{X}_1 = \mathrm{AVG}(\boldsymbol{X}) \tag{2.10}$$

$$\boldsymbol{Y}_1 = \mathrm{ReLU}(\mathrm{Linear}(\boldsymbol{X}_1)) \tag{2.11}$$

$$\boldsymbol{Y}_2 = \mathrm{ReLU}(\mathrm{Linear}(\boldsymbol{X}_1)) \tag{2.12}$$

$$\boldsymbol{Y}_3 = \mathrm{ReLU}(\mathrm{Linear}(\boldsymbol{X}_1))) \tag{2.13}$$

$$\boldsymbol{Y}_4 = \mathrm{ReLU}(\mathrm{Linear}(\boldsymbol{X}_1)) \tag{2.14}$$

$$\boldsymbol{Y} = \mathrm{Concat}(\boldsymbol{Y}_1, \boldsymbol{Y}_2, \boldsymbol{Y}_3, \boldsymbol{Y}_4) \tag{2.15}$$

其中，AVG( )表示全局平均池化，Linear( )表示全连接层，ReLU( )表示激活函数，Concat( )表示通道连接。融合特征 $\boldsymbol{Y}$ 经过全连接层和 Sigmoid 激活函数后生成权重矩阵，将权重矩阵与输入特征图相乘，得到加权特征图并给到检测头处理。如图 3-3 所示，SA 检测头增强网络的空间感知能力，提高对特征分类与定位的准确性。

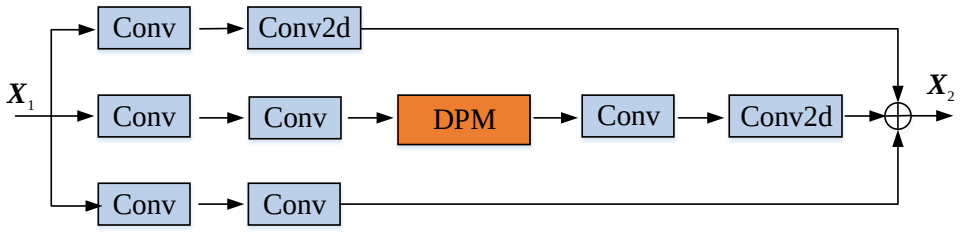


图 3-3 SA 检测头

3.2.2 显示视觉中心模块

本章引入了显示视觉中心模块，并行学习输入图像的局部角区域，其结构如图 3-4 所示。特征图 $\boldsymbol{X}$ 分别经轻量级多层感知机模块(Lightweight Multilayer Perceptron, LMLP)和可学习视觉中心模块(Location Visual Center, LVC)特征信息处理，然后将特征图在通道维度进行拼接，融合特征使用 1×1 卷积聚合特征信息，其过程如下：

$$\boldsymbol{X}_1 = \mathrm{Concat}\left(\mathrm{LVC}\left(\boldsymbol{X}\right), \mathrm{LMLP}\left(\boldsymbol{X}\right)\right) \tag{2.16}$$

$$\boldsymbol{X}_2 = \mathrm{Conv}^{1\times1}\left(\boldsymbol{X}_1\right) \tag{2.17}$$

其中，LVC( )表示可学习视觉中心处理特征信息，LMLP( )表示多层感知机处理特征信息，Concat( )表示特征图在通道维度进行拼接，$\mathrm{Conv}^{1\times1}$( )表示 1×1 卷积。

EVC 模块使用轻量级 MLP 来捕捉全局长程依赖，并利用 LVC 增强网络对通道信息的学习能力，通过编码和重标定机制使得每个通道能够根据其重要性进行调整，根据每个通道的重要性生成一个缩放因子，然后通过这个因子对原始输入进行通道级加权，从而使得网络能自适应地关注更加局部特征，不仅具有捕捉全局长程依赖的能力，还可以高效地获得全方位的、具有判别力的特征表示。

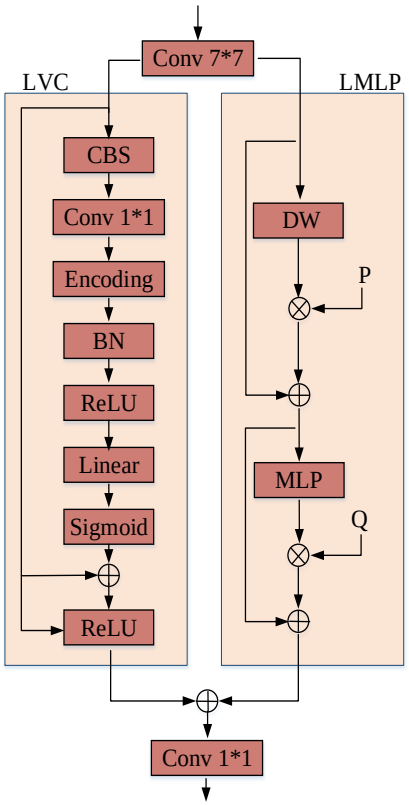


图 3-4 EVC 模块

3.3 实验设置与结果分析

3.3.1 实验设置

实验环境配置使用显卡为 NVIDIA GeForce RTX 2080Ti (11GB)、CPU 为四核 Intel(R) Xeon(R) Silver 4110CPU@2.10GHz、CUDA11.8 和 Pytorch2.0.1。在训练中，图像预处理成像素大小为 640×640，batchsize 设置为 8，使用马赛克数据增强方式，如图 3-5 所示，将训练图像随机缩放，多张图像拼接成一张大图作为新的训练数据。采用随机梯度下降(Stochastic Gradient Descent, SGD)的优化方式，优化器动量为 0.937。训练共 300 个 epoch，每个 epoch 使用 Complete-IoU 损失

函数计算损失，并记录当前的检测准确率、召回率和平均精度。

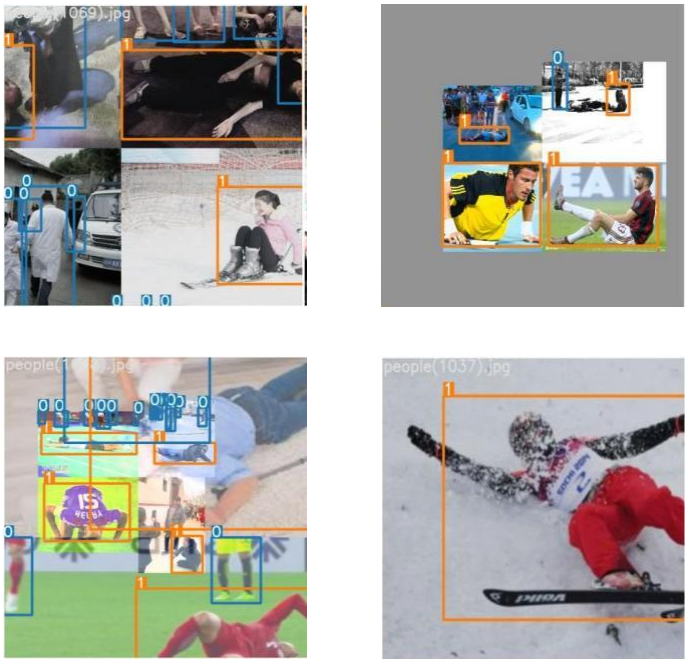


图 3-5 数据增强效果图

3.3.2 实验过程

本章对 YOLOv7-tiny 网络模型进行了优化，以提高对姿势识别检测的准确性。为了验证各项优化方法的有效性，本章在相同的实验条件下，在摔倒姿势数据集上进行了消融实验。实验 1 中引入了 SA 模块，实验 2 中同时引入 SER 模块和 EVC 模块，对比在测试集上的平均检测精度(mAP)和计算量(GFlops)，并对实验结果进行了详细的分析。

表 3-2 消融实验结果

实验	SA	EVC	mAP/%	GFlops/ Gflops
1	-	-	74.2	13.2
2	√	-	76.1	39.3
3	√	√	76.6	42.6

由表 3-2 中的实验结果对比可以看出，实验 2 中引入 SA 模块后，改进网络的检测精度比基线网络提高了 1.9%，说明改进网络关注重要区域的特征位置，提高了高级语义信息的利用；实验 3 引入 EVC 模块，相较基线网络检测精度提

高了 2.4%，说明改进网络聚焦局部信息，提高了信息交互能力。

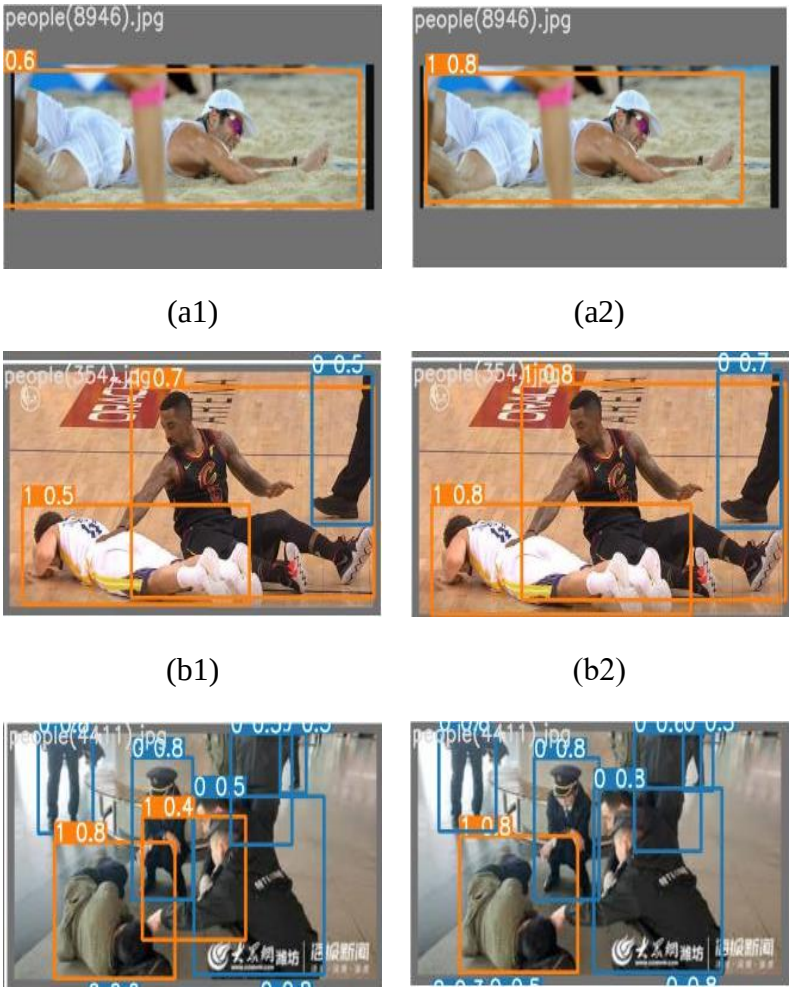


图 3-6 检测效果对比图

图 3-6 展示了检测效果。由图 a(1)和 a(2)的检测效果对比可以明显看出，改进网络对摔倒检测准确性提高，并且在人姿态信息不全的情况也可以做出准确判断。从图 3-6 的(b1)和(b2)的对比中可以直观地看出，低光照情况下，改进网络提高了对相似人体姿势的区分能力，能在摔倒姿势与半蹲、站立姿势中做出准确判断，避免了将非摔倒行为误认为摔倒而造成误检的可能性。

3.3.3 与其他方法对比

为了验证本章所提出的方法检测的有效性，对比试验基于人体摔倒姿势识

别数据集，将本章所提方法与其他目标检测网络模型进行对比。实验将平均精度(mAP)、参数量(Parameters)和计算量(GFLOPs)作为评价指标，列出了检测结果，如表 3-3 所示。

表 3-3 不同算法的对比实验结果

模型	mAP/%	Parameters/MB	GFLOPS/Gflops
YOLOv3	76.1	61.52	155.3
YOLOv5s	75.4	7.23	16.5
YOLOv7-tiny	74.2	6.02	13.2
YOLOv8s	75.5	11.17	28.8
SSD	69.9	26.15	62.7
本章算法	76.6	14.39	42.6

从表 3-3 中不同算法的实验结果对比可以看出，对人体姿势的识别过程中，本章提出的改进算法和其他网络相比，在检测精度上具有一定的优势。相较于 YOLOv7-tiny 网络，本章提出的算法优化了姿势特征信息的提取和处理方式，虽然网络复杂度和参数量增加，但是平均检测精度提高了 2.4%；相较于单阶段网络 YOLOv3 和 SSD，本章提出的改进模型在拥有较低的运算情况下，检测的准确度分别提高了 0.5%和 6.7%，检测精度和运行效率等方面更具有综合优势，可以满足检测的高效性需求。

3.4 本章小结

为了有效利用多尺度特征信息提高目标检测准确性，本章提出了一种基于 YOLOv7-tiny 改进的人体摔倒检测算法。首先，利用深度感知模块提升算法对重要信息的捕获能力，增强了检测头对于多尺度语义信息的处理能力；然后，在 head 部分引入显示视觉中心模块，提高算法对小尺度的高级语义信息处理。在人体摔倒数据集上，本章的改进算法对目标检测的平均检测精度达到了 76.6%，提高了 2.4%，但是检测速率有所下降，实时性降低，对此，后续章节中继续优化模型。

第 4 章 基于重参数化的自适应空间特征融合算法

4.1 引言

摔倒可能导致严重的健康问题，尤其是对于老年人、残疾人或患有慢性疾病的人群，如果能够及时发现摔倒情况，可以降低摔倒带来的危险。第三章引入深度感知检测头和视觉感知模块虽然一定程度上提高了轻量网络的检测能力，但是检测效率有所降低，为了解决这个问题，本章利用高效重参数化模块(Efficient Re-parameterization Feature Extraction Module, ERFEM)[92]提高网络检测效率，提升局部跨通道信息交互能力，增强浅层网络对人体姿势的局部信息的处理；然后，引入自适应空间特征融合结构(Adaptively Spatial Feature Fusion, ASFF)[93]，捕获人体姿态的全局信息，提高网络深层特征信息的融合能力；最后，将回归损失函数 CIoU[94]替换为 Shape-IoU[95]，提高网络的置信度。

4.2 算法原理

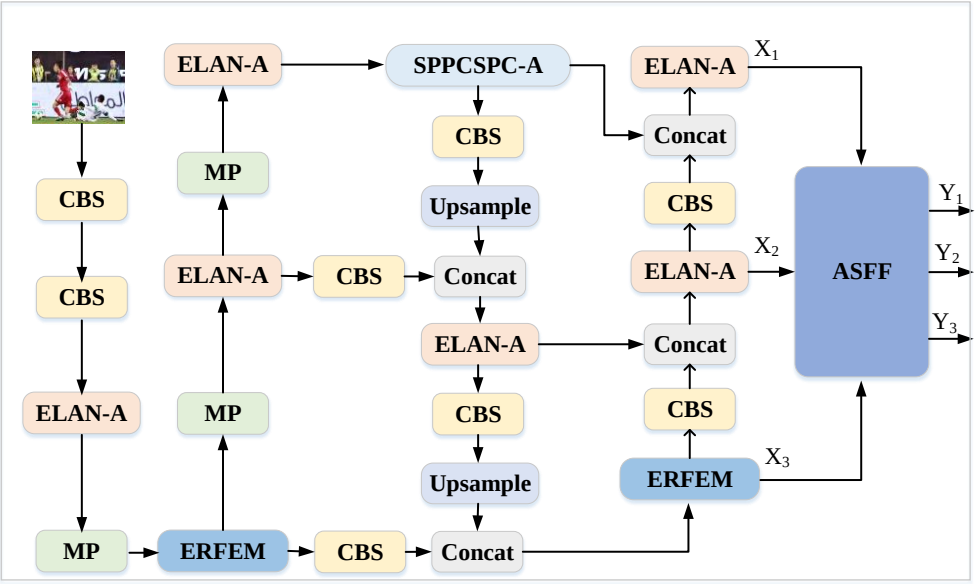


图 4-1 改进后的 yolov7-tiny 网络

本章基于 YOLOv7-tiny 对摔倒检测模型进行了改进。首先，通过引入高效的重参数化特征提取模块来增强特征金字塔结构中的浅层网络，使其对关键信息更

加敏感，提高网络对摔倒行为检测的效率；然后，利用自适应空间特征融合(ASFF)结构来提升深层网络中特征信息的融合利用，加强模型对全局信息的表征能力；最后，通过引入 Shape-IoU 损失函数，进一步提高了边界框回归的效率，增强了网络模型在检测摔倒行为时的精准性。YOLOv7-tiny 的改进后的网络结构如图 4-1 所示，其中 ELAN-A 表示轻量级 ELAN 结构，CBS 表示卷积层。

4.2.1 高效重参化特征提取模块

有效地利用局部特征信息可以增强轻量级网络在公共场合中的检测能力，在复杂的公共环境中，由于遮挡、视角变化等因素，人体的姿态可能会出现部分姿态特征缺失的情况。为提高对重要特征信息的关注提高摔倒行为的判别的可信度，本章引入了基于重参数幽灵(Re-parameterization Ghost, RepGhost)和高效通道注意力结构的高效重参化特征提取模块，有效提高特征提取能力。高效重参数化特征提取模块结构如图 4-2 所示。

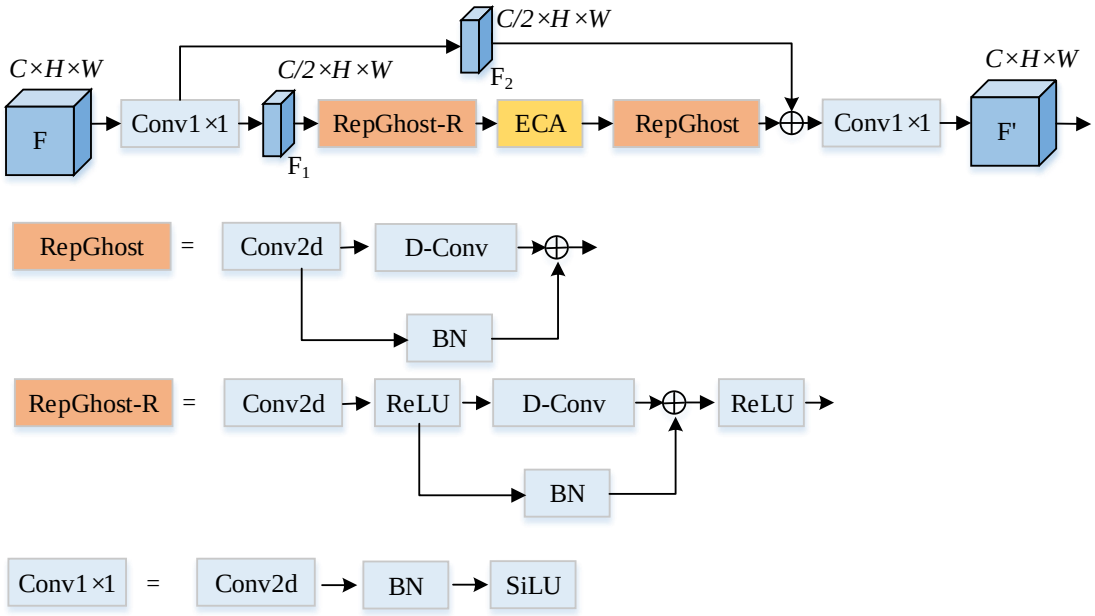


图 4-2 高效重参数特征提取模块

如图 4-2 所示，为了减少梯度信息冗余，特征图 $\boldsymbol{F} \in \boldsymbol{R}^{C\times H\times W}$ 经过卷积核大小为 1 的卷积层后，将特征信息分离为 $\boldsymbol{F}_1$、$\boldsymbol{F}_2 \in \boldsymbol{R}^{C\times H\times W}$。RepGhost 先由卷积核大小为 1 的卷积处理输入特征，再利用深度卷积(Deep Convolution, D-Conv)并进行批量归一化(Batchnorm, BN)，最后在权重空间进行特征融合，实现重参数化。

RepGhost 提高局部特征信息的表达，同时提升网络运行速度，具体过程如公式所示：

$$\boldsymbol{X}_1 = \mathrm{Conv}^{1\times1}\left(\boldsymbol{X}\right) \tag{4.1}$$

$$\boldsymbol{X}_2 = \mathrm{BN}\left(X_1\right) + \mathrm{D}^{3\times3}\left(\boldsymbol{X}_1\right) \tag{4.2}$$

其中，$\mathrm{Conv}^{1\times1}$ 是卷积核大小为 1 的卷积，BN 是批量归一化，$\mathrm{D}^{3\times3}$ 是深度卷积。

为提升特征提取模块对通道特征的敏感新，特征 $\boldsymbol{F}_1$ 在由添加了 ReLU 激活函数的重参数化模块后，ECA 模块对其进行处理。ECA 通道注意力层以提升模型对通道特征的敏感性，如图 4-3 所示。

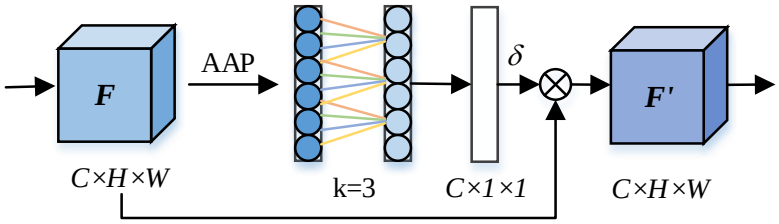


图 4-3 高效通道注意力模块

对每个通道全局平均池化，然后考虑每个通道及其 k 个邻居来捕获局部跨通道交互信息。最后，特征经过 RepGhost 模块的输出与特征图 $\boldsymbol{F}_2$ 进行尺度相加操作，并将结果输出。采用特征重参化的方式对特征进行隐式复用，提高浅层网络的特征信息捕获能力，并且嵌入通道注意力，定位重要局部特征，减少无关信息，提高模型的检测效率。

4.2.2 自适应空间特征融合结构

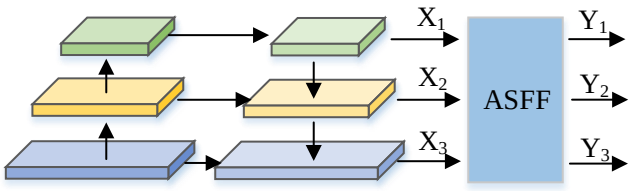


图 4-4 自适应空间特征融合结构

在基于特征金字塔的检测网络中，不同特征尺度之间的不一致性会显著降低检测网络的识别精度，这种不一致性主要来源于两个方面：一是不同层级特征图之间的分辨率差异导致的空间信息不一致；二是低层特征与高层特征之间语义信息的差异。为了解决这个问题，本章引入自适应空间特征融合方法，学

习空间过滤冲突信息抑制不一致性，提高网络的特征尺度不变性并增强模型对高级语义信息的处理能力。ASFF 的结构如图 4-4 所示。

ASFF 首先使用卷积核大小为 3 的卷积层丰富特征并提高通道维度，然后通过卷积层自适应地学习不同尺度的特征映射融合的空间权重，并利用学习权重参数将不同层次的特征集成在一起，如公式(4.3)所示：

$$\boldsymbol{Y}_n = \alpha_n \boldsymbol{X}_1 + \beta_n \boldsymbol{X}_2 + \gamma_n \boldsymbol{X}_3 \tag{4.3}$$

式中，$\boldsymbol{X}_1$、$\boldsymbol{X}_2$ 和 $\boldsymbol{X}_3$ 为不同层的特征向量，当三个不同层次的特征映射到第 n 层时，网络自适应学习空间重要性权重 α_n、β_n 和 γ_n。ASFF 结构将其他层次的特征整合到某一层次中，并调整到相同的分辨率，并对其他层次的特征信息进行空间过滤，保留有用的信息并用于组合，从而提高对深层信息的融合处理，人的姿态特征识别检测的准确性提高。

4.2.3 Shape-IoU 损失函数

YOLOv7-tiny 的回归框损失 CIoU 利用边界盒的相对位置和形状计算损失，虽然一定程度上提高了检测精度，但是 CIoU 忽略了边界盒形状和尺寸对 bounding box 回归的关系。本章引入 Shape-IoU 优化距离损失和形状损失，关注边界框本身的形状和尺寸，优化对重叠或不重叠区域、中心点距离、高度偏差等相关因素的计算，加速网络收敛，提高检测的精准度。如图 4-5 所示，A 是真值边界框，B 是预测边界框，真值边界框和预测边界框的中心点坐标分别为(x^A,y^B）和(x^B，y^B)，真值边界框的水平方向长度为 w^A，垂直方向为 h^A，预测边界框的水平方向为 w^B，垂直方向为 h^B。

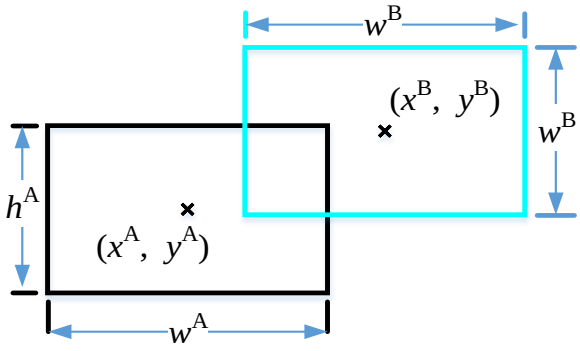


图 4-5 Shape-IoU 损失函数

Shape-IoU 定义如下：

$$L = 1 - IoU + Dis\tan ce - 0.5 \times \Omega \tag{4.4}$$

其中，IoU 是真值边界框和预测边界框的交并比，如公式(4.5)所示：

$$IoU = \frac{|A \cap B|}{|A \cup B|} \tag{4.5}$$

Distance 为距离损失的计算，过程如公式(4.6)~公式(4.8)所示：

$$w_1 = \frac{2 \times \left(w^A\right)^{scale}}{\left(w^A\right)^{scale} + \left(h^A\right)^{scale}} \tag{4.6}$$

$$h_1 = \frac{2 \times \left(h^A\right)^{scale}}{\left(w^A\right)^{scale} + \left(h^A\right)^{scale}} \tag{4.7}$$

$$Dis\tan ce = \frac{h_1 \times \left(x^B - x^A\right)^2}{c^2} + \frac{w_1 \times \left(y^B - y^A\right)^2}{c^2} \tag{4.8}$$

式中，w_1 和 h_1 是距离损失在水平方向和垂直方向的权重系数，*scale* 和 *c* 为可设置的参数。

Ω 为形状损失，定义如下：

$$w_2 = h_1 \times \frac{\left|w^B - w^A\right|}{\max\left(w^A, w^B\right)} \tag{4.9}$$

$$h_2 = w_1 \times \frac{\left|h^B - h^A\right|}{\max\left(h^A, h^B\right)} \tag{4.10}$$

$$\Omega = \left(1 - e^{-w_2}\right)^4 + \left(1 - e^{-h_2}\right)^4 \tag{4.11}$$

式中，w_2 和 h_2 是形状损失在水平方向和垂直方向的权重系数。

4.3 实验设置与结果分析

4.3.1 实验设置

本章实验环境配置为显卡 NVIDIA GeForce RTX 2080 Ti(11GB)、CPU 为 4 核 Intel(R) Xeon(R) Platinum 8255C CPU@2.50GHz、CUDA11.1 和 Pytorch1.9.0。在训练中，先将图像预处理成 640×640 大小，batchsize 设置为 8，采用随机梯度下

降（Stochastic Gradient Descent, SGD）的优化方式，优化器动量为 0.937。YOLOv7-tiny 的损失函数由分类损失、回归框损失和目标检测损失 3 部分构成，实验初始状态的回归框损失设置为 CIoU 损失函数。实验过程共训练 300 个 epoch，记录当前的检测 precison、recall 和 mAP。

表 4-1 实验环境设置

类别	型号
System	Windows10
CPU	Intel(R) Xeon(R) Platinum 8255C CPU@2.50GHz
GPU	NVIDIA GeForce RTX 2080Ti(11GB)
Framework	Pytorch1.9.0
Language	Python

4.3.2 实验过程

本章对 YOLOv7-tiny 网络模型进行了优化，以提高对姿势识别检测的准确性。为了验证各项优化方法的有效性，本章在相同的实验条件下，在摔倒姿势数据集上进行了消融实验。通过分别引入三个不同的结构，并使用不同的组合方式，对比了在测试集上的检测精度，并对实验结果进行了详细的分析。首先，实验 1 对基线网络 YOLOv7-tiny 进行测试，以此作为基准，评估后续各项改进措施的效果。然后实验 2 引入 RepGhost 模块，旨在提高网络的特征提取能力，同时实验 3 引入融合了 ECA 注意力的 ERFEM，通过增强通道注意力机制来提高特征提取的质量，特别是针对人体姿势的关键特征；实验 4 改进深层特征融合网络，引入 ASFF 模块，优化检测头以提高特征融合的质量和效率，减少信息冲突，增强深层信息的有效利用；实验 5 将回归函数 CIoU 替换为 Shape-IoU，并设置损失函数部分的比例因子 scale 为 0.7，c 为 2，优化了损失函数，使网络能够更好地处理形状和尺度的变化；实验 6 将 RepGhost 和 ECA 注意力机制以及 ASFF 模块同时添加至模型来评估这些改进措施共同作用的效果；实验 7 为最终整合了所有上述提到的优化方法的本章的检测模型。其中，AP_1 表示对人体其他姿势的识别精度，AP_2 表示对人体摔倒姿势的识别精度。

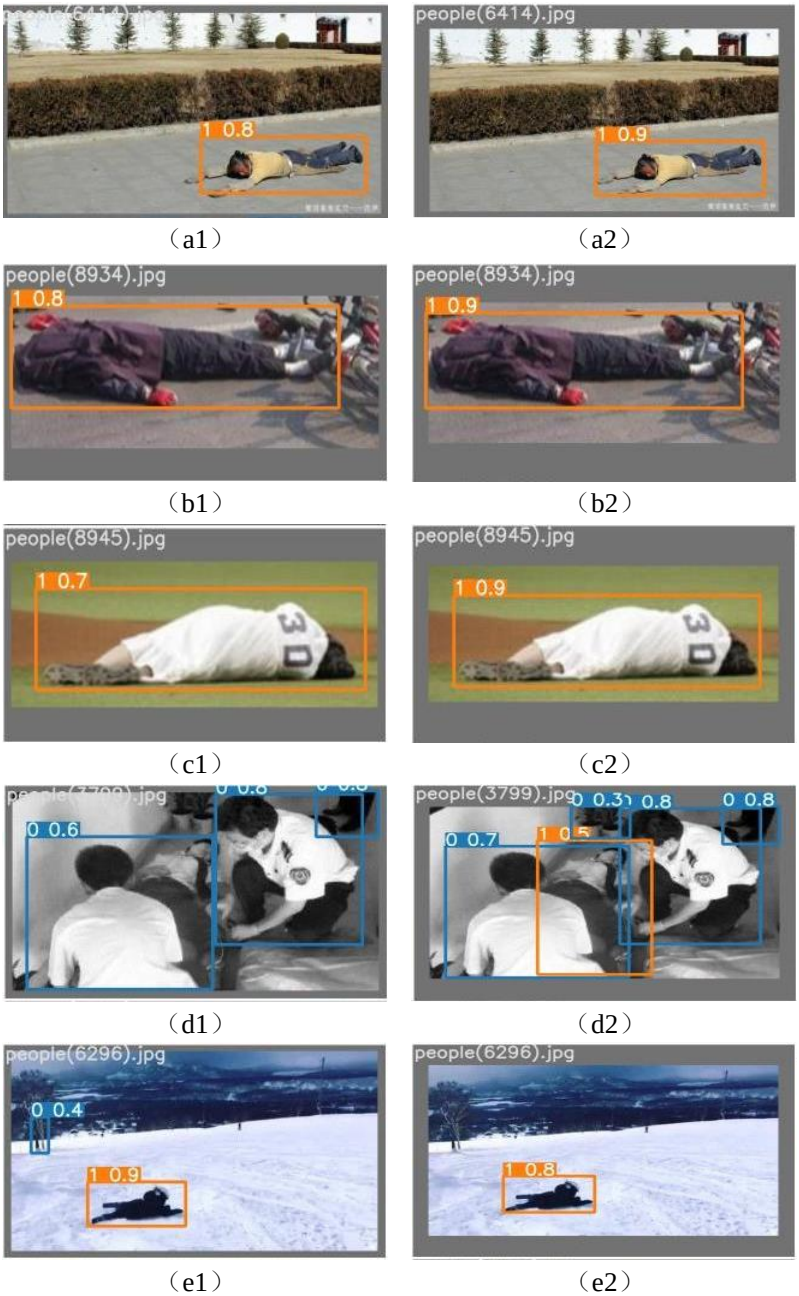
表 4-2 消融实验结果

实验	Repghost	ERFEM	ASFF	Shape-IoU	AP_1/%	AP_2/%	mAP/%
1	-	-	-	-	64.8	83.6	74.2
2	√				67.3	83.8	75.6
3	√	√			67.6	84.1	75.9
4			√		66.1	85.5	75.8
5				√	64.6	84.3	74.5
6	√	√	√		67.7	85.8	76.8
7	√	√	√	√	68.1	86.2	77.2

根据表 4-2 中的实验结果对比可以看出，实验 2 引入了 RepGhost 模块之后，平均检测精度比基线网络提高了 1.4%，对于姿势识别的敏感性得到了显著提高，并且实验 3 中 ERFEM 融合了通道注意力机制，平均检测精度较基线网络提高了 1.7%，说明改进网络进一步增强浅层网络的特征提取能力，提高了网络信息交互能力，进而提高了姿势识别的准确性；实验 4 中引入了 ASFF 之后，特征尺度不变性得到了提高，平均检测精度较基线提高了 1.6%，说明改进网络减少了特征融合网络中的信息冲突，促进了深层信息的有效利用；实验 5 中引入了 Shape-IoU，通过关注回归框损失的形状损失和距离损失，平均检测精度提高了 0.3%，摔倒姿势的识别精度提高了 0.7%，说明 Shape-IoU 提升了网络面对检测目标形状和尺度变化较大情况下的信息处理能力。实验 6、实验 7 依次添加改进部分，最终，优化网络检测的平均精度达到了 77.2%，较基线网络提高了 3%，其中，摔倒姿势的检测精度为 86.2%，提高了 2.6%，其他姿势的检测精度为 68.1%，提高了 3.3%，说明本章的改进模型，有效地提高了姿势识别的能力，进一步提升摔倒姿势的检测能力。

图 4-7 展示了其在网络检测效果对比的情况，其中不同的类别使用了不同的颜色框进行标注，摔倒的检测对象使用黄色检测框，其他姿势的检测对象则使用蓝色检测框。图(a1)~(e1)是 YOLOv7-tiny 的检测效果，图(a2)~(e2)是改进网络的检测效果。从图 4-7 的(a)~(c)检测效果对比可以明显看出，在多种不同的场景中，改进后的网络相较于基线网络显著增强了对人的摔倒特征的识别能力，提高了对

不同角度（例如前趴、仰卧、侧摔）摔倒对象检测的准确度。图 4-7 的(d)直观地表明，在颜色信息缺失、摔倒目标被遮挡的低光照情况下，改进网络有效降低了网络对复杂场景中检测目标漏检的可能性，对实际应用有重大意义。从图 4-7 的(e)的效果对比可以看出，改进网络即使在背景复杂情况下，也能不受周围环境的影响，准确地识别出摔倒的行为。



（a1）　（a2）

（b1）　（b2）

（c1）　（c2）

（d1）　（d2）

（e1）　（e2）

图 4-6 检测效果对比图

4.3.3 与其他方法对比

为了验证本章所提出的方法检测的有效性，对比试验基于人体摔倒姿势识别数据集，将本章所提方法与其他目标检测网络模型进行对比。实验将平均精度(mAP)、准确率(P)、召回率(R)、计算量(GFLOPs)和帧率(FPS)作为评价指标，列出了检测结果，如表 4-3 所示。

表 4-3 不同网络的对比实验结果

模型	mAP/%	P/%	R/%	Parameters/MB	GFLOPS/Gflops	FPS/fps
YOLOv3	76.1	78.4	72.4	61.52	155.3	135
YOLOv5s	75.4	77.5	71.4	7.23	16.5	277
YOLOv7-tiny	74.2	72.3	71.5	6.02	13.2	370.4
YOLOv8s	75.5	75.8	69.6	11.17	28.8	212
SSD	69.9	-	-	26.15	62.7	122
本章算法	77.2	79.6	70.7	29.35	14.5	294

从表 4-3 中不同算法的实验结果对比可以看出，对人体姿势的识别过程中，本章提出的改进算法和其他轻量级网络相比，在检测精度上取得了明显的优势。相较于单阶段网络 YOLOv3 和 YOLOv5s，本章提出的改进模型在拥有较低的运算情况下，检测的准确度分别提高了 1.1%和 1.8%；相较于 YOLOv7-tiny 网络，本章提出的算法优化了姿势特征信息的提取和处理方式，虽然网络复杂度和参数量略增，但是平均检测精度提高了 3%，并且对比 SSD 网络，检测精度和运行效率等方面更具有综合优势，可以满足检测的高效性需求。

4.4 本章小结

本章提出了一种基于 YOLOv7-tiny 改进的人体摔倒检测算法，通过融合了 ECA 和 RepGhost 的高效重参化特征提取模块，提高了网络对人体姿势的局部细节特征的捕获能力；再利用 ASFF 加权处理全局信息，提高了融合特征的尺度不变性，实现了上下文信息的有效利用；同时，通过引入 Shape-IoU 优化回归框，提高了网络对人体姿势的形状和尺度不一的检测能力，增强摔倒姿势识别能力；

最终，三者共同增强了检测的准确性。在人体摔倒数据集上，本章的改进算法对目标检测准确度有着明显提升，实现了实时准确检测，并且通过与其他轻量级网络对比，充分验证了改进算法的有效性。

第 5 章 联合信息增强和特征融合的改进 YOLOv7-tiny 算法

5.1 引言

复杂场景下，遮挡、光照不足、检测目标尺寸小、干扰目标多等因素影响人体姿势的实时检测能力，论文第三章和第四章对单阶段目标检测网络的深层信息进行处理，虽然一定程度上提高了多尺度特征网络的特征融合效率，实现检测准确性的提高，但是忽略了浅层特征信息的丢失问题。为了进一步满足摔倒检测实时性和准确性的需求，本章节在主干网络引入对比感知全局信息增强模块，增强对特征信息的追踪能力和敏感性，提取到更为丰富的语义信息，有效降低目标识别漏检的可能性。然后，为了充分利用不同层级的特征，引入融合坐标注意力[96](Coordinate Attention, CA)的密集坐标特征融合结构，提高不同层之间的信息交互，缓解信道信息聚合不充分的问题，提高网络对摔倒目标的检测精度。

5.2 算法原理

本章基于 YOLOv7-tiny 对行人摔倒模型进行改进。首先在主干网络的 L0、L1、L2 层后面添加 CAGIE 模块，当网络下采样后，提高网络对关键区域信息的敏感度，使得网络对人体姿势的学习更加充分，增强主干网络的表征能力；然后在特征融合部分引入 DCA 模块，弱化卷积操作带来的归纳偏置，并且聚焦关键特征的位置信息，提高网络信息利用和检测能力。改进后的网络结构如图 5-1 所示。

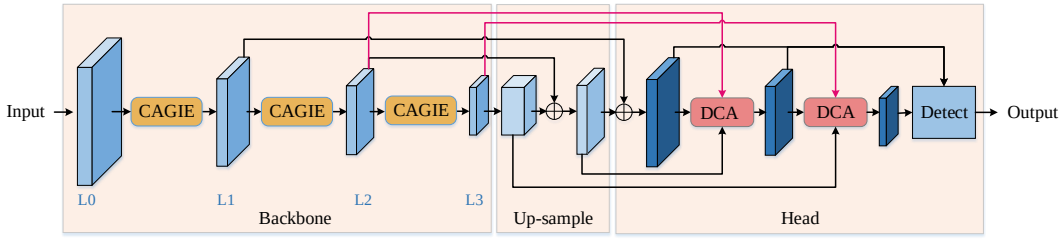
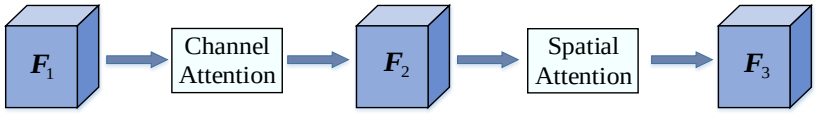


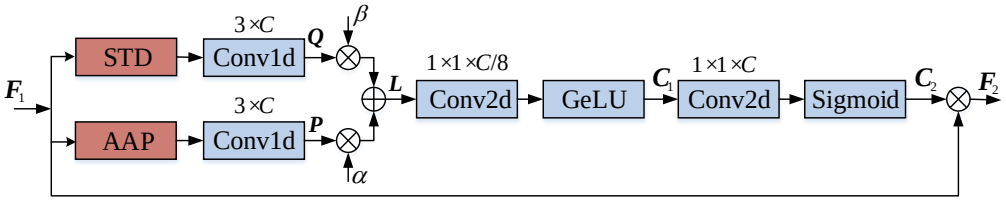
图 5-1 Yolov7-tiny 改进后的结构

5.2.1 对比感知全局信息增强模块

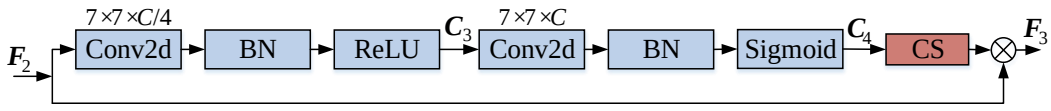
对不同感受野的特征信息进行处理和凝练，可以有效改善网络的检测性能。本章在主干网络引入对比感知全局信息增强模块(Contrast Aware Global Information Enhancement Module, CAGIE)，在通过最大池化层聚焦特征图后，学习特征图在不同感受野下的上下文信息，放大全局维度交互特征。CAGIE 模块按照通道-空间注意力的顺序设计了通道注意力和空间注意力两个注意力结构，在三维通道、空间宽度和空间高度上捕获重要特征以提高模型的识别精度。CAGIE 结构如图 5-2 所示。



(a)CAGIE 模块



(b)通道注意力模块



(c)空间注意力模块

图 5-2 对比感知全局信息增强模块

首先，为了重新整合与分配特征的通道权重，对不同感受野的高级语义特征 $\boldsymbol{F}_1 \in \mathbf{R}^{C\times H\times W}$（其中，C 为通道数，H 为图像高度，W 为图像宽度）使用通道注意力机制，如图 5-2(b)所示。通过自适应平均池化 AAP($\boldsymbol{F}_1$) 获取通道信息，利用一维卷积进行通道间信息交互得到全局平均池化信息 $\boldsymbol{P} \in \mathbf{R}^{C\times 1\times 1}$。为了获取细粒度上下文信息，选择标准差 STD($\boldsymbol{F}_1$) 处理特征，并使用卷积核大小为 3 的一维卷积得到全局对比度信息 $\boldsymbol{Q} \in \mathbf{R}^{C\times 1\times 1}$。为了获取聚合通道信息，使用可学习的参数调节 $\boldsymbol{P}$、$\boldsymbol{Q}$ 的比重：

$$\boldsymbol{L} = \alpha \cdot \boldsymbol{P} + \beta \cdot \boldsymbol{Q} \tag{5.1}$$

其中，$\boldsymbol{L}$是融合信息，$\boldsymbol{L}\in\mathbf{R}^{C\times1\times1}$；$\alpha$、$\beta$是初始值为 1 的可学习参数。

然后，再通过两个 1×1 的卷积组成的瓶颈结构，丰富特征信息并过滤无关信息，并使用 GeLU 激活函数增加非线性和 Sigmoid 激活函数得到权重。利用权重对特征图进行处理，计算公式为：

$$\boldsymbol{C}_1=\mathrm{GeLU}\left(\mathrm{Conv}_{1\times1}(\boldsymbol{L})\right) \tag{5.2}$$

$$\boldsymbol{C}_2=\sigma\left(\mathrm{Conv}_{1\times1}(\boldsymbol{C}_1)\right) \tag{5.3}$$

$$\boldsymbol{F}_2=\boldsymbol{C}_2\otimes\boldsymbol{F}_1 \tag{5.4}$$

其中，σ表示 Sigmoid 激活函数；$\otimes$是逐元素相乘；$\mathrm{Conv}_{1\times1}$是卷积核大小为 1 的二维卷积；$\boldsymbol{C}_1$和$\boldsymbol{C}_2$是使用不同激活函数的卷积层的特征向量；$\boldsymbol{F}_1$是输入通道注意力的特征图，$\boldsymbol{F}_2$是特征图$\boldsymbol{F}_1$经过通道注意力处理后的特征图。

特征的通道信息经过通道注意力后得到适当调整，再通过空间注意力提高特征的空间信息表征，增强全局信息的交互能力和网络的表达能力。如图 5-2(c)所示，空间注意力先用第一个卷积层在通道维度上对特征映射$\boldsymbol{F}_2\in\mathbf{R}^{C\times H\times W}$进行整合和压缩，将特征维度下降为原先的 4 倍，然后聚集上下文信息，再用第二个卷积层使维度扩大至 4 倍，特征从$\mathbf{R}^{C/4\times H\times W}$映射为$\mathbf{R}^{C\times H\times W}$。每个卷积层为减小信息离散度进行数据正则化，7×7 的大核卷积层使网络更聚焦于检测目标，提高空间信息聚集。最后，通过对特征图通道信息重组(Channel Shuffle, CS)提高特征相关性，空间注意力的计算如下：

$$\boldsymbol{C}_3=\mathrm{ReLU}\left(\mathrm{BN}\left(\mathrm{Conv}_{7\times7}(\boldsymbol{F}_2)\right)\right) \tag{5.5}$$

$$\boldsymbol{C}_4=\sigma\left(\mathrm{BN}\left(\mathrm{Conv}_{7\times7}(\boldsymbol{C}_3)\right)\right) \tag{5.6}$$

$$\boldsymbol{F}_3=\mathrm{CS}\left(\boldsymbol{C}_4\right)\otimes\boldsymbol{F}_2 \tag{5.7}$$

其中，$\mathrm{Conv}_{7\times7}$是卷积核大小为 7 的卷积；BN(·)是数据批归一化处理；$\boldsymbol{C}_3$和$\boldsymbol{C}_4$是使用不同激活函数的卷积层的特征向量；CS(·)是对特征图的通道信息重组；$\boldsymbol{F}_3$是将特征图$\boldsymbol{F}_2$经过空间注意力处理后的特征图。

5.2.2 密集坐标注意力特征融合结构

YOLOv7-tiny 的特征融合网络将 FPN 顶层的强语义信息与 PAFPN 自下而上

的强定位信息相结合，提高信息的利用程度。但是，密集人群中人的姿态信息易受周围环境影响而缺失，卷积神经网络在经过多次的卷积操作后，部分高级特征信息的丢失降低了关键姿势信息的捕获能力，导致漏检小尺度的群体目标。为了减少高级语义信息的丢失数量，提高网络的检测精度，本章提出了融合坐标注意力(Coordinate Attention, CA)机制的密集坐标注意力特征融合模块(Dense Coordinate Attention Feature Fusion Module, DCA)，通过增强网络多层级之间的特征交互，加深网络对重要特征的关注度。DCA 结构如图 5-3 所示。

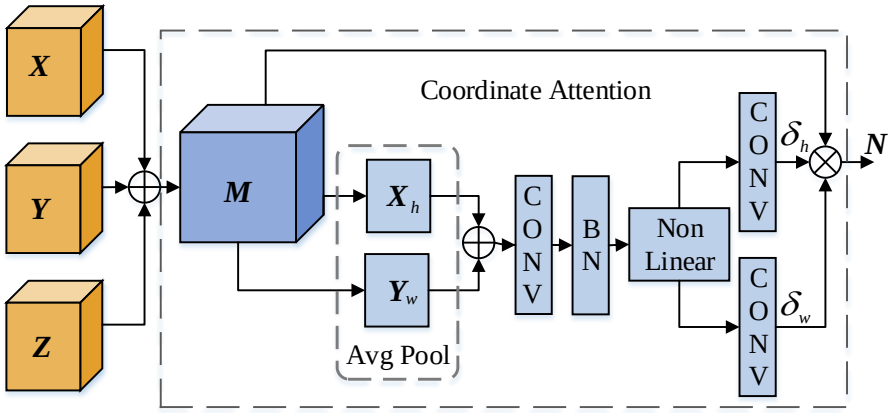


图 5-3 密集坐标注意力特征融合模块

图 5-3 中三个不同层的特征图 $\boldsymbol{X},\boldsymbol{Y},\boldsymbol{Z}\in\mathbf{R}^{C\times H\times W}$，$\boldsymbol{X}$ 是浅层语义特征图，$\boldsymbol{Y}$ 是由 $\boldsymbol{X}$ 深度卷积后上采样得到的中层特征图，$\boldsymbol{Z}$ 是深层语义特征图。首先，将三个不同层的特征图在通道维度上进行连接后得到融合特征图 $\boldsymbol{M}$，提高了网络多层级特征之间的信息交互能力。其次，利用 CA 模块处理密集连接的融合特征图 $\boldsymbol{M}$，分别在水平和垂直两个方向上进行全局特征池化以减少网络运行的计算量，图中 $\boldsymbol{X}_h$ 是水平方向的特征向量，$\boldsymbol{Y}_w$ 是垂直方向的特征向量。然后，将这两个特征向量拼接起来得到全局特征信息 $C(\boldsymbol{X}_h\oplus\boldsymbol{Y}_w)$，选用 1×1 卷积聚合特征信息并由 BN 层进行归一化处理后，对特征信息进行编码。为了获得细粒度特征坐标信息，利用 1×1 的卷积提取局部特征，并使用 Sigmoid 激活函数获得特征图在高度和宽度上的注意力权重。最后，在融合特征图上进行乘法加权计算，生成具有坐标感知的特征图。特征在不同位置具有不同的注意力水平，计算公式为：

$$\boldsymbol{N}=\boldsymbol{M}\times\delta_h\times\delta_w \tag{5.8}$$

其中，δ_h 是高度方向的注意力权重；δ_w 是宽度方向的注意力权重；$\boldsymbol{M}$ 是不同层级的融合特征图；$\boldsymbol{N}$ 是 $\boldsymbol{M}$ 经 CA 模块后的具有权重信息的特征图。

DCA 模块采用特征重用的思想，并注重考虑融合特征图的空间位置信息，增

强模型网络对人体姿势的检测关注，在推理过程中更关注特征权值高的区域，抑制特征重用时不重要的背景信息造成的干扰，建立具有精确位置信息的长期依赖关系。

5.3 实验设置与结果分析

5.3.1 实验设置

实验环境配置如表 5-1 所示，显卡为 NVIDIA GeForce RTX 2080Ti (11GB)、CPU 为四核 Intel(R) Xeon(R) Silver 4110CPU@2.10 GHz 、CUDA11.8 和 Pytorch2.0.1。在训练中，图像预处理成像素大小为 640×640，batchsize 设置为 8，使用马赛克数据增强方式，随机拼接、旋转、缩放图片。采用随机梯度下降(Stochastic Gradient Descent, SGD)的优化方式，优化器动量为 0.937。训练共 300 个 epoch，每个 epoch 使用 Complete-IoU 损失函数计算损失，并记录当前的检测准确率、召回率和平均精度。

表 5-1 实验环境配置

类别	型号
System	Windows10
CPU	Intel® Xeon(R) Silver 4110CPU@2.10 GHz
GPU	NVIDIA GeForce RTX 2080Ti(11 GB)
Framework	Pytorch2.0.1
Language	Python

5.3.2 实验数据集

本章使用 2019 年中国华录杯数据湖算法大赛定向算法（人体摔倒姿势识别）提供的图片资源和百度 AIStudio 建立的开源跌倒检测数据集 Fall detection Database，同时，为验证模型的通用性，选择使用学生课堂行为(Student Classroom Behavior, SCB)数据集和 PASCAL VOC 数据集来评估本章改进算法，如表 5-2 所示。学生课堂行为数据集中 SCB1 包含举手类别，共 4001 张图片（训练集 3171 张，测试集 830 张）；SCB2 包含举手、读书、写字三个类别，共 4266 张图片（训

练集 3418 张，测试集 848 张）。PASCAL VOC 数据集中包含行人等 20 个类别的检测目标，其中训练集由 PASCAL VOC 2007 训练集和 PASCAL VOC 2012 训练集组成，并且实验在 PASCAL VOC2007 测试集上进行测试。

表 5-2 数据集分布

数据集	训练集/张	测试集/张	验证集/张
Fall	6225	779	779
SCB1	3171	830	-
SCB2	3418	848	-
PASCAL VOC	16551	4952	-

5.3.3 实验评价指标

为了更好地评估改进模型与 YOLOv7-tiny 及其他模型的性能，在相同的实验条件下，本章引入准确率(Precision)、召回率(Recall)、平均精度(Mean Average Precision, mAP)、计算量(Giga Floating-point Operations Per Second, GFLOPs)、帧率（Frames Per Seconds, FPS）和参数量作为评价指标。

$$P = \frac{N_{\mathrm{TP}}}{N_{\mathrm{TP}} + N_{\mathrm{FP}}} \times 100\% \tag{5.9}$$

$$R = \frac{N_{\mathrm{TP}}}{N_{\mathrm{TP}} + N_{\mathrm{FN}}} \times 100\% \tag{5.10}$$

$$P_{\mathrm{mAP}} = \frac{\sum_{i=1}^{k} P_{\mathrm{AP}_i}}{k} \tag{5.11}$$

其中，P 表示准确率；R 表示召回率；N_{TP} 是真阳性(True Positives)，表示目标检测正确；N_{FP} 是假阳性(False Positives)，表示目标检测错误，将负样本的目标误判为正样本的目标；N_{FN} 是假阴性(False Negatives)，表示目标检测错误，将正样本的目标错误判断为负样本；P_{AP_i} 表示第 i 类目标的检测精度；P_{mAP} 表示 k 类目标的检测精度平均值。

5.3.4 实验过程

本章在人体摔倒姿势识别数据集上进行了一系列消融实验，以证明改进的YOLOv7-tiny网络各个模块的有效性。在每个对比实验中，设置相同的实验参数。本章利用平均精度、计算量、准确率和召回率分析每个模块的功能，如表5-3所示。表中M1是CAGIE模块；M2是DCA结构；为了验证CAGIE模块的有效性，将DCA结构中的CA注意力替换为CAGIE模块，即M3；√表示网络中加入该模块，×表示不加入该模块。

表5-3 改进YOLOv7-tiny的消融实验

模型	M1	M2	M3	P_{mAP}/%	计算量/Gflops	Precision/%	Recall/%
1	×	×	×	73.3	13.2	77.8	64.6
2	√	×	×	75.4	14.2	76.5	69.2
3	√	√	×	77.0	14.5	75.5	70.8
4	√	×	√	76.5	22.6	77.2	69.8

由表5-3可以看出，CAGIE模块虽然增加了网络的运算，但是提升了网络捕获信息的能力，增强了网络的特征表达能力。表5-4给出两种类别的检测精度，其中0代表非摔倒的检测目标，1代表摔倒的检测目标。实验1是在YOLOv7-tiny主干网络中添加CAGIE模块，实验2同时引入CAGIE模块和DCA结构。

表5-4 跌倒检测数据集上的检测结果

模型	0-P_{AP}/%	1-P_{AP}/%	均值P_{mAP}/%
YOLOv7-tiny	62.4	84.1	73.3
实验1	65.2	85.6	75.4
实验2	67.9	86.0	77.0

从表5-4的mAP结果对比可知，实验1添加了CAGIE模块，平均精度比YOLO7-tiny增加了2.1%，非摔倒目标的检测精度增加了2.8%，摔倒目标的检测精度增加了1.5%，表明CAGIE模块通过平均池化和标准差的方式获取了更细粒度的特征信息，使用大核卷积提高了网络的信息感受能力，增强了特征提取网络的信息捕获能力；实验2中改进网络的平均检测精度达到了77.0%，相较于原始

网络，平均精度提高了 3.7%，其中非摔倒目标的检测精度提高了 5.5%，摔倒目标的检测精度提高了 1.9%，表明改进的 YOLOv7-tiny 网络在细化特征信息后充分利用上下文信息，降低人群目标检测漏检的可能，提升了对非摔倒目标和摔倒目标的识别能力，增强了网络整体的检测准确性。

图 5-4 摔倒检测效果对比

图 5-4 展示了 YOLOv7-tiny 和改进的 YOLOv7-tiny 在摔倒检测数据集上的效果对比，图中用不同颜色框出检测目标，非摔倒人群使用蓝色检测框，摔倒人群用黄色检测框。由图 5-4 可以看出，相较于原网络，改进网络的特征捕获能力大大提高，降低了网络对不同场景下检测目标漏检的可能性，能够检测出更多密集的、小尺度的摔倒目标，增强了对被遮挡以及特征缺失的检测目标的识别能力，提高了摔倒姿态的检测精度，改善了在低光照环境中的检测性能。

5.3.5 与其他方法对比

为了验证在增强主干网络提取特征能力后，引入 CA 注意力的特征融合结构对上下文信息利用的有效性，本章进行了对比实验。将 CA 注意力分别替换为高效特征注意力(Efficient Channel Attention, ECA)和 CBAM 注意力，表 5-5 给出融合不同注意力的实验效果。

表 5-5 不同注意力特征融合的对比实验结果

模块	P_{mAP}/%	计算量/Gflops
ECA	75.8	14.4
CBAM	75.0	14.6
CA	77.0	14.5

由表 5-5 可知，与 CBAM 模块相比，CA 和 ECA 计算需求略小，并且都将融合特征的通道信息进行加权处理，进一步提高网络的信息检索；与 ECA 模块相比，CA 更注重融合特征图的空间位置，提高多层级语义信息的特征提取，充分学习权值信息，提升网络对人体姿势信息的学习能力和检测精度，验证了本章改进方法的有效性。

为了验证本章算法对摔倒姿势检测的准确性，对比实验基于该数据集的摔倒类别与其他算法进行比较。实验将不同算法的平均精度、帧率、计算量和参数量作为评价指标，表 5-6 给出不同算法的检测结果。

表 5-6 针对摔倒类别的不同算法对比实验

模型	P_{mAP}/%	计算量/Gflops	帧率/fps	参数量/MB
改进的 FCOS[97]	82.2	193.0	-	3.76
SSD	80.1	-	113.1	-
YOLOv3	80.3	155.3	135	61.52
YOLOv5s	80.7	16.5	277	7.23
YOLOv8s	85.5	28.8	212	11.17
YOLOv7-tiny	84.1	13.2	370	6.02
本章算法	**86.0**	**14.5**	**294**	**6.83**

由表 5-6 可以看出，与单阶段网络 YOLOv3、YOLOv5s 和 YOLOv8s 相比，

本章算法在拥有较低的运算量情况下，摔倒目标的检测精度分别提高了 5.7%、5.3%和 0.5%；与 FCOS 的改进算法[94]相比，本章算法对摔倒目标的检测精度提高了 3.8%；与 YOLOv7-tiny 网络相比，本章算法增加了 CAGIE 模块和 DCA 结构，虽然网络参数量略增，检测速度略微降低，但摔倒类别的检测精度提高了 1.9%。对比其他经典网络，本章算法在检测精度和运行效率等方面更具有综合优势，可以满足检测的高效性需求。

为验证本章算法具有良好的通用能力，在学生课堂行为数据集和 PASCAL VOC 上进行实验。表 5-7 给出 YOLOv7-tiny 和本章算法在学生课堂行为上的检测结果。

表 5-7 学生课堂行为数据集上的检测结果

数据集	模型	举手 P_{AP} /%	读书 P_{AP} /%	写字 P_{AP} /%	均值 P_{mAP} /%
SCB1	YOLOv7-tiny	77.8	-	-	77.8
	本章算法	78.4	-	-	78.4
SCB2	YOLOv7-tiny	93.2	88.0	90.7	90.6
	本章算法	93.3	88.7	91.2	91.1

从表 5-7 可以看出，本章算法在 SCB1 和 SCB2 数据集上平均检测精度为 78.4%和 91.1%，与 YOLOv7-tiny 对比，分别提高了 0.6%和 0.5%。在 SCB1 和 SCB2 数据集上的学生课堂行为检测效果如图 5-5 和图 5-6 所示。

图 5-5 SCB1 数据集上的检测效果对比

图 5-6 SCB2 数据集上的检测效果对比

对比图 5-5 和图 5-6 的检测效果，本章算法能够准确地识别举手行为，在密集的小尺寸检测目标中可以判别出更多的举手行为，特征学习能力得到提高，降低了被漏检的可能性，并且本章算法能够准确区分不同的人体姿势行为，具有一定的通用性。表 5-8 给出不同算法在 VOC 2007 测试集上得到的平均检测精度。

表 5-8 不同算法在 VOC2007 测试集上的平均检测精度

模型	主干网络	P_{mAP}/%
Faster R-CNN	ResNet-101	76.4
FCOS	ResNet-50	78.7
改进的 CenterNet[98]	ResNet-50	80.9
改进的 Faster R-CNN[99]	Res2Net-101	79.7
YOLO	DarkNet	63.4
YOLOv2	DarkNet53	78.6
YOLOv3	DarkNet53	80.3
改进的 YOLOv4-tiny[100]	-	72.1
YOLOv5s	CSPDarkNet53	79.1
YOLOv7-tiny	CSPDarkNet53	79.4
本章算法	CSPDarkNet53	**81.5**

从表 5-8 可以看出，本章算法在检测准确性上具有明显优势。相较于双阶段检测算法中的 Faster R-CNN，本章算法的检测精度提高了 5.1%；相较于单阶段目标检测算法的 YOLOv5s 和 YOLOv7-tiny，本章算法的检测精度分别提高了 2.4%和 2.1%；相较于其他改进算法，如改进的 Faster R-CNN 和改进的 YOLOv4-tiny，检测精度提高了 1.8%和 9.4%，证明本章算法具有良好的目标检测能力和普适性。

表 5-9 本章算法在 VOC 2007 测试集上各类别检测精度对比

类别	YOLOv7-tiny	Ours
Aeroplane	87.5	$89.8^{+2.3}$
Bicycle	88.5	$90.4^{+1.9}$
Bird	73.1	$76.2^{+3.1}$
Boat	67.5	$72.2^{+4.7}$
Bottle	70.4	$73.7^{+3.3}$
Bus	88.1	$90.0^{+1.9}$
Car	91.6	$92.4^{+0.8}$
Cat	85.0	$86.5^{+1.5}$
Chair	62.0	$64.5^{+2.5}$
Cow	84.2	$84.8^{+0.6}$
Dining table	73.9	73.3
Dog	81.4	$85.1^{+3.7}$
Horse	88.7	$89.3^{+0.6}$
Motorbike	87.6	$89.5^{+1.9}$
Person	**86.3**	$\mathbf{87.9^{+1.6}}$
Potted plant	52.3	$55.8^{+3.5}$
Sheep	81.8	$84.4^{+2.6}$
Sofa	72.2	$75.3^{+3.1}$
Train	86.9	$88.2^{+1.3}$
TV monitor	79.3	$81.4^{+2.1}$
mAP/%	**79.4**	$\mathbf{81.5^{+2.1}}$

从表 5-9 中检测结果可以看出，相对于基线网络，改进后的网络对大部分类别的目标具有显著提高，尤其对小目标如 bird（鸟）、bottle（瓶子）、potted plant（盆栽）这些小目标的检测准确性提升显著，分别有 3.1%、3.3%、3.5%的提高，并且，对行人的检测精度提高了 1.6%。

图 5-7 展示了本章改进算法对 PASCAL VOC2007 数据集部分类别的检测效果。从图 5-7 可以看出，改进算法可以有效检测出目标物体，并且在低光照、遮挡、密集的环境中仍能很好的检测出物体。

图 5-7 PASCAL VOC 2007 检测效果图

5.4 本章小结

本章提出了一种基于 YOLOv7-tiny 改进的人体摔倒检测算法，通过对比感知全局信息增强模块对比特征通道信息，提高了主干网络感知特征信息的能力，加强了改进网络对重要特征的学习能力；再利用密集坐标注意力特征融合结构聚合不同层级的语义信息，加权处理融合特征，实现了上下文信息的有效利用；最终，两者共同增强了网络对摔倒行为的检测准确性。在人体摔倒数据集上，改进算法对目标检测准确度有着明显提升，实现了实时准确检测，并且通过与其他方法对比，充分验证了改进算法的有效性，同时，在学生课堂行为数据和 PASCAL VOC 数据集的实验结果表明，改进算法具有一定的泛化能力。

第 6 章 总结与展望

6.1 工作总结

论文主要对基于图像的单目标检测网络进行研究，通过多尺度特征融合和信息增强等方法，提高单阶段目标检测网络的检测准确性，具体如下：

（1）针对轻量网络 YOLOv7-tiny 特征提取能力不足的问题，本文提出了一种深度感知的多尺度目标检测算法，其分别引入了显示视觉中心模块和空间感知检测头。其中，显示视觉中心模块通过编码和重标定机制提高了网络对局部特征的关注，增强网络局部特征信息的提取能力；空间感知检测头通过深度感知模块增强网络对特征信息的定位能力，实现了深度信息的有效利用。在摔倒数据集上的进行验证，实验结果说明所提方法提高了特征提取能力，深、浅层特征信息有效表达提升了检测的准确度。

（2）针对深层网络中存在特征信息丢失的问题，本文分别引入了对比感知全局信息增强模块、密集坐标特征融合结构和自适应空间特征融合结构。对比感知全局信息增强模块嵌入到主干网络中，提高了网络对高分辨率重要特征信息的敏感性，从而提取到更为丰富的语义信息，降低了目标识别过程中漏检的可能性；密集坐标特征融合结构和自适应空间特征融合结构分别利用跨层加权和自适应学习的特征融合方式，提高网络的特征尺度不变性和多层级之间的特征交互，充分利用不同层级的特征，有效解决特征丢失的问题。

（3）针对 YOLOv7-tiny 在处理深层特征信息时检测效率下降的问题，本文引入高效重参数化模块优化网络结构，不仅增强网路局部跨通道信息的交互能力，浅层网络能够更好地处理人体姿态的局部细节信息，还提高了检测效率。

（4）训练策略优化。本文引入 Shape-IoU 优化回归损失函数，进一步提高网络的置信度，同时，使用马赛克数据增强方式，训练图像随机缩放，并将多张图像拼接成一张大图，生成新的训练数据，避免数据过拟合。

在摔倒数据集上，本文提出的改进算法在目标检测准确度有着明显提升，具有实时性，并且本文的第五章内容中，在学生课堂行为数据集和 PASCAL VOC

2007 数据集上进行实验与分析，验证了所提出的算法具有一定的泛化性。

6.2 工作展望

本论文主要基于图像的单阶段目标检测算法研究，提出不同方式提高检测准确性，一定程度上提高了目标检测能力。论文虽然在第二章中介绍了融合了点云信息的图像目标检测方法，但是并没有对其进行深入研究。点云信息准确的三维位置信息可以缓解了图像目标检测定位的不准确性问题，但在三维场景到二维图像的映射过程中缺乏有效的几何约束，并且对点云的序列信息没有很好的利用。另外，基于图像的目标检测可以应用在视频跟踪中，利用连续帧信息，可能进一步降低检测误判，多模态信息融合的目标检测是未来目标检测边缘部署的有效方式。在本文的研究基础上，还有很多值得探索的研究方向，例如针对于小尺度的检测目标的检测准确性仍有待提高。论文以摔倒姿势为主要研究姿势，但人体姿态复杂，并且姿态信息在遮挡等情况下，姿态信息具有一定的模糊性，如何利用轻量化的算法对各种姿态给出准确预测具有一定的实际应用价值。

参考文献

[1] LeCun Y, Bengio Y, Hinton G. Deep learning[J]. Nature, 2015, 521: 436-444.

[2] 王梓祺，李阳，张睿，等. 小样本 SAR 图像分类方法综述[J]. 中国图像图形学报，2024，29(07)：1902-1920.

[3] Zhou X Y, Gong W, Fu W L, et al. Application of deep learning in object detection[C]//2017 IEEE/ACIS 16th International Conference on Computer and Information Science, Wuhan, China, IEEE, 2017: 631-634.

[4] Taghanaki S A, Abhishek K, Cohen J P, et al. Deep semantic segmentation of natural and medical images: A review[J]. Artificial Intelligence Review, 2021, 54: 137-178.

[5] 黄雯珂，滕飞，王子丹，等. 基于深度学习的图像分割综述[J]. 计算机科学，2024，51(02)：107-116.

[6] 马双双，王佳，曹少中，等. 基于深度学习的二维人体姿态估计算法综述[J]. 计算机系统应用，2022，31(10)：36-43.

[7] 王仕宸，黄凯，陈志刚，等. 深度学习的三维人体姿态估计综述[J]. 计算机科学与探索，2023，17(01)：74-87.

[8] Dhiman C, Vishwakarma D K. A review of state-of-the-art techniques for abnormal human activity recognition[J]. Engineering Applications of Artificial Intelligence, 2019, 77: 21-45.

[9] 王凤随，邵凯丽，杨海燕. 联合信息增强和特征融合的人体摔倒检测算法[J]. 中国惯性技术学报，2024，32(08)：771-778.

[10] Thanoon M A, Zulkifley M A, Mohd Z M A A, et al. A review of deep learning techniques for lung cancer screening and diagnosis based on CT images[J]. Diagnostics, 2023, 13: 2617.

[11] 贺晓松，胡川丽，赵江. 基于深度学习的非小细胞肺癌检测[J]. 临床放射学杂志，2025，44(02)：265-273.

[12] 杨永胜，邓淼磊，李磊，等. 基于深度学习的行人重识别综述[J]. 计算机工程

与应用，2022，58(09)：51-66.

[13]王坤，项琦鑫. 改进 Yolov4 的车辆弱目标检测算法[J]. 中国惯性技术学报，2023，31(08)：797-805.

[14]刘菲，钟延芬，邱佳伟. 基于改进 YOLOv5s 的轻量化交通标志识别检测算法[J]. 激光与光电子学进展，2024，61(24)：102-114.

[15]吴瑞林，葛泉波，刘华平. 基于 YOLOX 的类增量印刷电路板缺陷检测方法[J]. 智能系统学报，2024，19(04)：1061-1070.

[16]左敏，纪慧卓，苏礼君，等. 人工智能在食品安全中的最新应用及进展[J]. 中国食品学报，2024，24(10)：1-13.

[17]谢俊，李玉萍，左飞飞，等. 基于机器视觉的孔类零件尺寸在线检测[J]. 电子测量技术，2021，44(02)：93-98.

[18]许玉格，钟铭，吴宗泽，等. 基于深度学习的纹理布匹瑕疵检测方法[J]. 自动化学报，2023，49(04)：857-871.

[19]Papageorgiou C P, Oren M, Poggio T A general framework for object detection[C]//Sixth International Conference on Computer Vision. New York: IEEE, 1998: 555-562.

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

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

[22]Ojala T, Pietikainen M, Harwood D. Performance evaluation of texture measures with classification based on Kullback discrimination of distributions[C]// Proceedings of 12th International Conference on Pattern Recognition. New York: IEEE, 1994: 582-585.

[23]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. New York: IEEE, 2016: 779-788.

[24] REDMON J, FARHADI A. YOLO9000: Better, faster, stronger[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2017: 7263- 7271.

[25] Redmon J, Farhadi A. YOLOv3: An incremental improvement[J]. ArXiv preprint arXiv:1804.02767, 2018.

[26] Bochkovskiy A, Wang C Y, Liao H Y M. YOLOv4: Optimal speed and accuracy of object detection[J]. ArXiv preprint arXiv:2004.10934, 2020.

[27] 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.

[28] Li C, Li L, JIANG H, et al. YOLOv6: A single-stage object detection framework for industrial applications[J]. ArXiv preprint arXiv:2209.02976, 2022.

[29] 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:2207.02696, 2022.

[30] Wang C Y, Yeh I H, Liao H Y M. YOLOv9: Learning what you want to learn using programmable gradient information[J]. Arxiv preprint arXiv:2402.13616, 2024.

[31] WANG A, CHEN H, LIU L, et al. YOLOv10: RealTime End-to-End Object Detection[J]. Arxiv preprint arXiv:2405.14458, 2024.

[32] Liu W, Anguelov D, Erhan D, et al. SSD: Single shot multibox detector[C]//European Conference on Computer Vision. Cham: Springer, 2016: 21-37.

[33] Fu, C-Y, Liu, W, Ranga, A, Tyagi, A, Berg, A C. DSSD: Deconvolutional single shot detector[J]. Arxiv preprint arXiv:1701.06659, 2017.

[34] Li Z, Zhou F. FSSD: Feature fusion single shot multibox detector[J]. ArXiv preprint arXiv:1712.00960, 2017.

[35] Zhai, S, Shang, D, Wang, S, Dong, S. DF-SSD: An improved SSD object detection algorithm based on DenseNet and feature fusion[J]. IEEE Access, 2020, 8: 24344-24357.

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

[37] Zhou X, Zhuo J, Krahenbuhl P. Bottom-up object detection by grouping extreme and center points[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2019: 850-859.

[38] 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. New York: IEEE, 2019: 9627-9636.

[39] 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. New York: IEEE, 2019: 6569-6578.

[40] Ren S Q, He K M, Girshick R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks[J]. Advances in Neural Information Processing Systems, 2015, 39(6): 1137-1149.

[41] He K M, Gkioxari G, Dollár P, et al. Mask R-CNN[C]//Proceedings of the IEEE International Conference on Computer Vision. New York: IEEE, 2017: 2961-2969.

[42] Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//2014 IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2014: 580-587.

[43] He K M, Zhang X Y, Ren S Q, 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.

[44] Girshick R. Fast R-CNN[C]//2015 IEEE International Conference on Computer Vision(ICCV). New York: IEEE, 2015: 1440-1448.

[45] Li Z M, Peng C, Yu G, et al. Light-head R-CNN: In defense of two-stage object detector[J]. Arxiv preprint arXiv:1711.07264, 2017.

[46] 李昌财，陈刚，侯作勋，等. 自动驾驶中的三维目标检测算法研究综述[J]. 中国图象图形学报，2024，29(11)：3238-3264.

[47]郭帅. 基于相机和激光雷达融合的目标检测算法研究[D]. 吉林大学，2024.

[48]Qi C R, Su H, Mo K, et al. PointNet: Deep learning on point sets for 3D classification and segmentation[C]//2017 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2017: 77-85.

[49]Qi C R, Yi L, Su H, et al. PointNet++: Deep hierarchical feature learning on point sets in a metric space[C]//Proceedings of the Advance in Neural Information Processing Systems(NeurIPS). Red Hook: Curran Associates Inc., 2017: 4-9.

[50]Zhou Y, Tuzel O. VoxelNet: End-to-end Learning for point cloud Based 3D object detection[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2018: 4490-4499.

[51]Yan Y, Mao Y X, Li B. SECOND: Sparsely embedded convolutional detection[J]. Sensors, 2018(10): 3337.

[52]Lang A H, Vora S, Caesar H, et al. PointPillars: Fast encoders for object detection from point clouds[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2019: 12689-12697.

[53]Qi C R, Liu W, Wu C X, et al. Frustum PointNets for 3D object detection from RGB-D data[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2018: 918-927.

[54]Ku J, Mozifian M, Lee J, et al. Joint 3D Proposal generation and object detection from view aggregation[C]//2018 IEEE/RSJ International Conference on Intelligent Robots and System(IROS). New York: IEEE, 2018: 1-8.

[55]Mohapatra S, Yogamani S K, Gotzig H, et al. BEVDetNet: Bird’s eye view LiDAR point cloud based real-time 3D object detection for autonomous driving[C]//2021 IEEE International Intelligent Transportation Systems Conference(ITSC). New York: IEEE, 2021: 2809-2815.

[56]Zhou J, Tan X, Shao Z, et al. FVNet: 3D front-view proposal generation for real-time object detection from point clouds[C]//2019 12th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics(CISP-

BMEI). New York: IEEE, 2019: 1-8.

[57] He C H, Zeng H, Huang J Q, et al. Structure aware single-stage 3D object detection from point cloud[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). New York: IEEE, 2020: 11870-11879.

[58] Yang Z T, Sun Y N, Liu S, et al. 3DSSD: Point-based 3D single stage object detector[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New York: IEEE, 2020: 11037-11045.

[59] Zheng W, Tang W L, Jiang L, et al. SE-SSD: Self-ensembling single-stage object detector from point cloud[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New York: IEEE, 2021: 14489-14498.

[60] Noh J, Lee S, Ham B. HVPR: Hybrid voxel-point representation for single-stage 3D object detection[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New York: IEEE, 2021: 14600-14609.

[61] Li J Y, Luo C X, Yang X D. PillarNeXt: Rethinking network designs for 3D object detection in LiDAR point clouds[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New York: IEEE, 2023: 17567-17576.

[62] Hu Y, Lu Y F, Xu R S, et al. Collaboration helps camera overtake LiDAR in 3D detection[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New York: IEEE, 2023: 9243-9252.

[63] Lin T Y, Dollár P, Girshick R, et al. Feature pyramid networks for object detection[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2017: 936-944.

[64] Liu S.; Qi L, Qin H F, et al. Path aggregation network for instance segmentation[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2018: 8759-8768.

[65] Ghiasi G, Lin T Y, Le Q V. NAS-FPN: Learning scalable feature pyramid architecture for object detection[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2019: 7029-7038.

[66]Tan M, Pang R, Le Q V. EfficientDet: Scalable and efficient object detection[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2020:10778-10787.

[67]Li Y H, Chen Y T, Wang N Y, et al. Scale-aware trident networks for object detection[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV). New York: IEEE, 2019: 6053-6062.

[68]郑嘉硕，张子木，唐声权，等. 基于双 YOLOv8 结构的多模态图像目标检测技术研究[J/OL]. 宇航计测技术，1-11[2025-02-27].

[69]罗一凯，何林远，马时平. 基于多尺度特征的多模态激光雷达增强算法[J]. 激光与光电子学进展，2024，61(18)：238-246.

[70]周牧，冉浩成，王勇，等. 基于激光雷达点云投影的多视图融合目标检测方法[J/OL]. 光学学报，1-16[2025-02-27].

[71]杨锁荣，杨洪朝，申富饶，等. 面向深度学习的图像数据增强综述[J/OL]. 软件学报，2025，36(03)：1390-1412.

[72]Yun S, Han D, Chun S, et al. CutMix: Regularization strategy to train strong classifiers with localizable features[C]//2019 IEEE/CVF International Conference. on Computer Vision. New York: IEEE, 2019, 6022-6031.

[73]Devries T, Taylor G W. Improved regularization of convolutional neural networks with cutout. Arxiv preprint arXiv:1708.04522, 2017.

[74]Zhang H Y, Cisse M, Dauphin Y N, et al. Mixup: Beyond empirical risk minimization. Arxiv preprint arXiv: 1710.09412, 2017.

[75]Goodfellow I J, Pouget-Abadie J, Mirza M, et al. Generative adversarial networks[J]. Communications of the ACM, 2020, 63(11): 139-144.

[76]任书玉，汪晓丁，林晖. 目标检测中注意力机制综述[J]. 计算机工程，2024，50(12)：16-32.

[77]Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate[J]. Computer Science, 2014.

[78]Hu J, Shen L, Sun G. Squeeze-and-excitation networks[C]//Proceedings of the

IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2018: 7132-7141.

[79] Woo S, Park J, Lee J, et al. CBAM: Convolutional block attention module[C]//Computer Vision-ECCV 2018. Cham: Springer, 2018: 11211.

[80] Carion N, Massa F, Synnaeve G, et al. End-to-end object detection with transformers[C]//Proceedings of European Conference on Computer Vision. Berlin: Springer International Publishing, 2020: 213-229.

[81] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems(NIPS'17). Red Hook: Curran Associates Inc, 2017: 6000-6010.

[82] Huang G, Liu Z, Van Der Maaten L, et al. Densely connected convolutional networks[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition(CVPR). New York: IEEE, 2017: 2261-2269.

[83] 徐世文，王姮，张华，等. 一种基于关键点的红外图像人体摔倒检测方法[J]. 红外技术，2021，43(10)：1003-1007.

[84] Tanwar R, Nandal N, Zamani M, et al. Pathway of trends and technologies in fall detection: a systematic review[C]//Healthcare. MDPI, 2022, 10(1): 172.

[85] Orejel Bustos A S, Tramontano M, Morone G, et al. Ambient assisted living systems for falls monitoring at home[J]. Expert review of medical devices, 2023, 20(10): 821-828.

[86] Zhang H B, Zhang Y X, Zhong B, et al. A comprehensive survey of vision-based human action recognition methods[J]. Sensors. 2019, 19(5):1005.

[87] Wang X Y, Ellul J, Azzopardi G. Elderly fall detection systems: a literature survey[J]. Frontiers in Robotics and AI, 2020, 7: 71.

[88] Jiang S L, Qi H Z, Zhang J, et al. A pilot study on falling-risk detection method based on postural perturbation evoked potential features[J]. Sensors, 2019, 19(24): 5554.

[89] Shao Y, Wang X Y, Song W J, et al. Feasibility of using floor vibration to detect

human falls[J]. International Journal of Environmental Research and Public Health, 2021, 18(1): 200.

[90] Hassan C A U, Karim F K, Abbas A, et al. A cost-effective fall-detection framework for the elderly using sensor-based technologies[J]. Sustainability, 2023, 15(5): 3982.

[91] Rastogi S, Singh J. Human fall detection and activity monitoring: a comparative analysis of vison-based methods for classification and detection techniques[J]. Soft Computing, 2022, 26(8): 3679-3701.

[92] Chen C, Guo Z, Zeng H, et al. Repghost: a hardware-efficient ghost module via re-parameterization[J]. Arxiv preprint arXiv:2211.06088, 2022.

[93] 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. New York: IEEE, 2021: 13029-13038.

[94] Zheng Z, Wang P, Ren D, et al. Enhancing geometric factors in model learning and inference for object detection and instance segmentation[J]. IEEE Transactions on Cybernetics, 2021, 52(8): 8574-8586.

[95] Zhang H, Zhang S. Shape-IoU: More accurate metric considering bounding box shape and scale[J]. Arxiv preprint arXiv:2312.17663, 2023.

[96] Hou Q, Zhou D, Feng J. Coordinate attention for efficient mobile network design[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2021: 13708-13717.

[97] 方可，刘蓉，魏驰宇，等. 复杂场景下的行人跌倒检测算法[J]. 计算机应用，2023，43(06)：1811-1817.

[98] 熊磊，王凤随，钱亚萍. 基于特征融合的自适应多尺度无锚框目标检测算法[J]. 电子测量与仪器学报，2022，36(11)：236-244.

[99] Wang S, Qu Z. Multiscale anchor box and optimized classification with faster R-CNN for object detection[J]. IET Image Processing, 2022, 17(5):1322-1333.

[100] Baces D, Oniga F. Single stage architecture for improved accuracy real-time

object detection on mobile devices[J]. Image and Vision Computing, 2023, 130: 104613.

致谢

快乐的时光总是一闪而过。三年时间给予我沉淀的机会。

有师如斯，庆幸之至。首先，我要特别感谢我的导师王凤随老师。回首往昔，王老师的严谨态度一直都鼓舞着我，鼓励开展自己感兴趣的课题。还记得研一那会儿，因为课设作业不认真，被王老师谈话，让我意识到自己学习态度上的问题，也让我对学术研究有了全新的认识，更加明白了何为责任和追求卓越。王老师不仅在学术上给予了我无尽的帮助，更重要的是他教会了我做事的态度，每一次对话、每一封邮件，无不是老师对学生的关心和支持，使我受益匪浅。

感谢我的家人们，他们是我最坚实的后盾。从小到大，家人一直鼓励我多读书，教育我要热爱学习。三孝口的新华书店是我童年的一角，在那里播种下了求知的种子。虽然在求学路上，我走的磕磕绊绊，但所幸，一直都是向上而行。在面对选择时，父母将很多的决定权交给了我，他们的信任和支持，是我在追求梦想道路上最大的支持。家人的无私奉献和默默付出，让我懂得了什么是爱，也教会了我如何去爱。

感谢我的朋友们，你们总是在身边默默支持着我，给予我力量。黄梦琦是我的快乐搭子，无论是跳舞还是走街串巷寻找美食，亦或是骑电动车去滨江看一场日落，她总是在我身边分享那些美好的瞬间。王路遥师姐，是我的心理导师，无论我有多少烦恼和困惑，她总是耐心倾听，抚平我的浮躁，又给予我建议。许月师姐、杨海燕师姐带我进入课题，储曦师兄会帮助我修改程序，可能一改就是好久，让我的研究生学习生活没有特别大的压力。卓文涛，我的涛姐，时常陪我锻炼身体，有她在宿舍，那必然是干净整洁。在研三上学期，每次回学校，杨苏朋都会去高铁站接我，一路上听我的碎碎念。姚鑫钧，我的宝藏男孩，在我对未来的规划感到迷茫，甚至有些畏缩时，鼓励我去做自己想做的事情，无论遇到什么困难，一切都会是光明的未来。很感谢我的朋友们，那些跳舞、打羽毛球的日子，无论想做什么，你们都在，好像所有事情都是那么轻而易举。记得当我说要减肥时，一群人浩浩荡荡在神山公园跑步，夕阳下的汗水都很快乐。这些珍贵的记忆，不仅是生活的调味剂，也是我坚持下去的动力源泉。

在这即将结束的时刻，心中满是对过去三年的感激之情。这段经历不仅丰富了我的专业知识，更教会了我如何做人、做事，感谢所有陪伴我走过这段旅程的人。未来的路还很长，但我相信，我一定能够走得更加坚定自信。愿我们在各自的人生旅途中都能绽放光彩，再次衷心感谢所有在我求学过程中给予我帮助和支持的人。最后，我要向百忙之中抽出宝贵时间参与论文评审和答辩的专家、老师们表达我最诚挚的谢意。

攻读硕士学位期间的研究成果

1.发表学术论文

[1] 王凤随，邵凯丽，杨海燕. 联合信息增强和特征融合的人体摔倒检测算法[J]. 中国惯性技术学报，2024，32(08)：771-778.

2.参与科研项目

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、国家自然科学基金项目(42206198)