

基于深度学习的目标检测算法轻量化研究综述

董甲东¹, 桑飞虎¹, 郭庆虎¹, 陈琳¹, 郑春香²⁺

1. 安庆师范大学 电子工程与智能制造学院, 安徽 安庆 246011

2. 安庆师范大学 计算机与信息学院, 安徽 安庆 246011

+ 通信作者 E-mail: zhcx@aqnu.edu.cn

摘要:随着深度学习的快速发展,基于深度学习的目标检测算法在多个领域得到广泛应用。然而,随着算法不断演进,一系列挑战也逐渐浮现:模型复杂度增加导致参数量和计算量急剧膨胀,进而使得运行速度下降,难以满足高实时性应用场景需求;算法对硬件性能高要求,限制其在移动设备和边缘计算等资源受限环境中高效部署,缩小其应用范围;训练成本显著上升,包括对大量计算资源和长时间训练需求,也阻碍模型快速迭代。为了应对这些挑战,目标检测算法轻量化研究应运而生。综述基于深度学习的目标检测算法轻量化研究最新进展。对目标检测任务及其评价指标进行概述,详细回顾目标检测算法发展历程及代表性模型。在此基础上,重点探讨目标检测算法的轻量化技术,包括:轻量级网络架构设计,降低模型计算复杂度和空间复杂度;轻量级卷积技术创新,在减少参数量和计算量的同时保持模型性能;深度学习模型压缩方法,优化模型结构降低存储需求。实现资源受限设备高效部署与实时推理。总结当前轻量化目标检测算法研究现状,并在多领域技术融合、硬件架构优化以及边缘设备部署等方面进行展望与思考。

关键词:深度学习;目标检测算法;模型轻量化

文献标志码:A **中图分类号:**TP391.41

Review of Lightweight Object Detection Algorithms Based on Deep Learning

DONG Jiadong¹, SANG Feihu¹, GUO Qinghu¹, CHEN Lin¹, ZHENG Chunxiang²⁺

1. School of Electronic Engineering and Intelligent Manufacturing, Anqing Normal University, Anqing, Anhui 246011, China

2. School of Computer and Information, Anqing Normal University, Anqing, Anhui 246011, China

Abstract: With the rapid development of deep learning, object detection algorithms based on deep learning have been widely used in many fields. However, with the continuous evolution of the algorithm, a series of challenges has gradually emerged: the increase in the complexity of the model leads to the rapid expansion of the number of parameters and the amount of computation, which in turn reduces the running speed and makes it difficult to meet the needs of high-real-time application scenarios. The high requirements of the algorithm on hardware performance limit its efficient deployment in resource-constrained environments such as mobile devices and edge computing, and narrow its application range. The significant rise in training costs, including the need for large computational resources and long training time, also prevents the rapid iteration of models. In order to deal with these challenges, the research of lightweight object detection algorithm comes into being. This paper aims to review the recent progress of lightweight object detection algorithms based on deep learning. In this paper, the task of object detection and its evaluation index are summarized, and then the

基金项目:国家自然科学基金(62205005);安徽省高校科研计划重大项目(2024AH040174)。

This work was supported by the National Natural Science Foundation of China (62205005), and the Anhui University Scientific Research Plan Major Project (2024AH040174).

收稿日期:2025-03-04 **修回日期:**2025-04-28

development history and representative model of object detection algorithm are reviewed in detail. On this basis, the paper focuses on the lightweight technology of object detection algorithm, including lightweight network architecture design to reduce model computational complexity and spatial complexity, lightweight convolutional technology innovation to reduce the number of parameters and computation while maintaining model performance, deep learning model compression method to optimize model structure to reduce storage requirements. Efficient deployment and real-time inference of resource-constrained devices are realized. Finally, this paper summarizes the current research status of lightweight object detection algorithm, and prospects and thinks in the aspects of multi-field technology integration, hardware architecture optimization and edge device deployment.

Key words: deep learning; object detection algorithm; model lightweight

数字化时代,图像和视频数据普遍使用,极大地促进人工智能和深度学习技术迅速增长,深度学习在图像处理领域的应用成为趋势。随着深度学习进步,目标检测技术也得到显著提升,并衍生出多种基于深度学习的目标检测算法,这些算法在自动驾驶、智能监控及工业检测等多个领域广泛应用。但是,深度学习模型通常设计复杂,有大量参数和计算需求,这使得其在资源有限的设备上运行效率低下。因此,轻量级的目标检测算法研究应运而生。

实现轻量化目标检测算法主要有以下途径。第一,目标检测通过轻量化神经网络架构的使用,减少计算复杂度和资源消耗,使其能够在资源受限的设备上高效运行。如He等^[1]通过主干网络替换,显著减少网络计算量和模型大小,提高检测速度。第二,目标检测通过轻量级卷积设计,减少参数量。如Simonyan等^[2]强调使用多个 3×3 卷积核堆叠代替单个大卷积核的优势,这种策略不仅减少参数量,还增加网络深度,从而提高模型性能。第三,目标检测在深度学习中利用模型压缩,通过减少模型参数量和存储空间来降低模型大小。如Chen等^[3]就是通过模型压缩方法,显著降低卷积神经网络(convolutional neural network, CNN)模型参数量和计算复杂度,提高模型效率和适应性。轻量化算法凭借这些优势在目标检测领域取得显著成果。

此前,有较多文献对目标检测算法进行总结分析,Mittal^[4]从模型压缩和加速,综述目标检测模型在边缘设备上应用的全面调查;董甲等^[5]从单阶段目标检测算法在金属表面缺陷检测的应用及优化方法进行综述,但未针对算法的轻量化深入调研;Kaur等^[6]将现有目标检测算法分为一阶段和两阶段进行综述,但未对模型设计进行深入分析。

本文首先介绍目标检测算法发展及轻量化的评价指标,重点总结轻量化目标检测算法关键技术,包括轻量级网络、轻量级卷积设计优化以及深度学习模型压缩方法。结合文献综述,本文从多学科技术融合、硬件架

构设计优化、边缘设备部署策略三个方面展望未来发展方向。

1 目标检测算法

目标检测是计算机视觉领域重要研究方向,旨在识别图像或视频中物体并标记其位置和分类。传统目标检测在处理复杂背景、大规模数据集和实时应用性能有限,手动提取特征耗时且通用性差。深度学习方法通过自动特征提取、端到端检测等方式学习更复杂的表示,显著提升了对各种变化和复杂场景的适应性,相较于传统方法具有更高的准确性和鲁棒性。

1.1 传统目标检测

传统目标检测需要进行多阶段处理,大致分为区域选择、特征提取和目标分类。此外,这些方法还需要预先大量人工标注数据,如AdaBoost^[7]算法、支持向量机^[8](support vector machine, SVM)、决策树^[9]等。

为生成候选区域,常用滑动窗口扫描整个图像。然而,该方法计算成本高,并且会导致大量不必要选择。因此,一些改进的算法如级联分类器^[10]被提出,其训练过程如图1所示。Ali等^[11]针对无人机交通监控,采用AdaBoost算法将多个弱分类器级联成强分类器,结合质心跟踪算法(centroid tracking)实现车辆跟踪,显著减少了计算成本并提升了检测速度。在AU-AIR数据集上检测精度达到95%,但依赖手工特征,受限于样本质量与复杂场景处理。

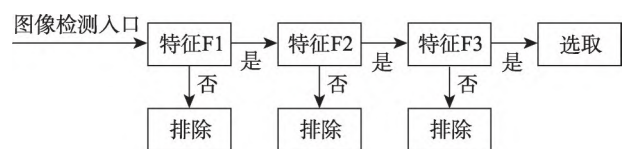


图1 级联分类器

Fig.1 Cascade classifier

定位对象后,进行特征提取,以提供鲁棒性表征。通常使用尺度不变特征变换^[12](scale-invariant feature

transform, SIFT)、方向梯度直方图^[13](histogram of oriented gradients, HOG)等手工设计的特征描述符。然而,手动构建出全面的特征描述符来正确识别各种物体非常困难,如SIFT特征提取流程如图2所示。Tang等^[14]为了解决高分辨率图像对齐计算复杂度高的问题,基于SIFT通过扩展极值点检测范围与非极大值抑制增强特征点稳定性,设计56维圆形描述符减少计算复杂度。在图像数据集城市场景上检测时间仅为原模型的0.10%,但处理无人机图像的实时性和复杂变换的能力不足。

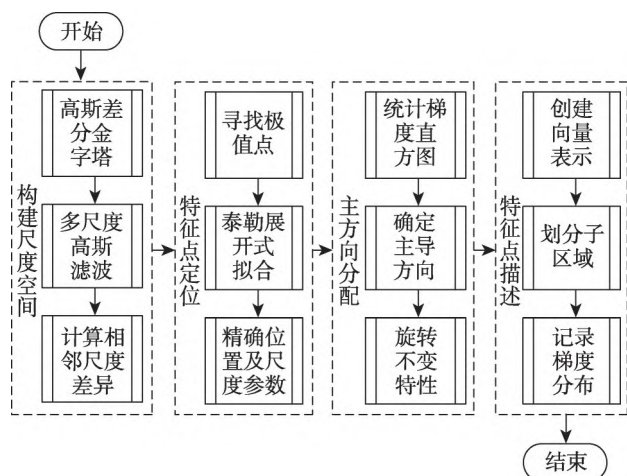


图2 SIFT特征提取流程

Fig.2 SIFT feature extraction process

在目标分类阶段,传统目标检测算法使用诸如决策树之类的机器学习模型,对提取的特征进行分类。其训练过程如图3所示。虽然决策树能处理多种数据类型,但容易过拟合且对噪声敏感。Barros等^[15]针对桥梁钢材的缺陷检测,提出基于模型的决策树算法。结合了振动监测数据和有限元模型,能够自动检测桥梁可能存在的损伤并定位损伤位置,从而提高检测效率,减少人为误差。但依赖高精度FE模型,特定桥梁需定制化设计泛化能力弱。

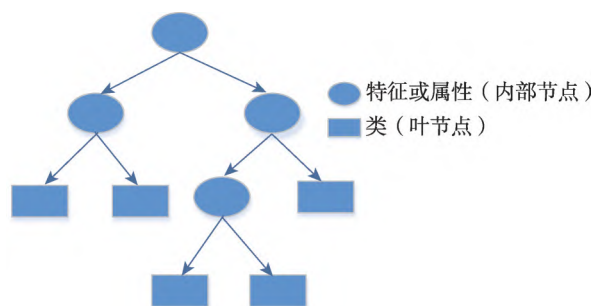


图3 决策树训练示意图

Fig.3 Decision tree training diagram

为提高模型性能,通过集成学习来实现优于传统目标检测的轻量化。如Zhang等^[16]改进的S-HOG特征提取方法,相较于传统HOG精度提升13.64%,模型大小仅为0.14 MB。Abbas等^[17]将ConvMixer架构与AdaBoost分类器结合,提出AFR-Conv算法。通过多模型特征融合策略以提高识别的鲁棒性,检测精度达到95.5%。然而在面对复杂场景和光照变化的目标检测任务时,无法达到深度学习算法轻量化的水平。

1.2 深度学习目标检测

基于深度学习的目标检测算法主要分为两类:第一类是两阶段目标检测算法,如R-CNN、Fast R-CNN和Faster R-CNN等。这种方法准确性得到提升,但牺牲一定速度;第二类是单阶段目标检测算法,如YOLO(you only look once)、SSD(solid state drive)和DETR(detection transformer)等。该方法在速度上更具优势,更适用于实时目标检测任务。

1.2.1 两阶段目标检测

随着Krizhevsky等^[18]提出AlexNet,并在ImageNet(ILSVRC)图像分类任务中成功应用,目标检测的发展开始逐渐走向深度学习领域。目标检测算法发展历程如图4所示。

两阶段目标检测算法常用选择性搜索(selective search, SS)或区域提议网络(region proposal network, RPN)等方法生成候选区域,然后对候选区域进行特征提取,使用分类器判断目标类别,并通过回归器调整边界框以精确定位目标。

这种方法的优点在于高精度和灵活性,适应不同任务需求。R-CNN系列作为典型代表,广泛应用于学术研究和工业检测领域。然而,该方法也存在不足:检测速度较慢,难以满足实时性要求;对硬件资源,如GPU(graphics processing unit)要求高。

尽管后续研究提出了改进算法,如NAS-FPN、Sparse RCNN和H2RBox-v2等如表1所示,但仍存在模型结构复杂、计算资源需求大等问题。

针对上述问题,杨扬等^[29]对特征金字塔网络(feature pyramid network, FPN)进行改进提出了ResFPN算法。通过设计轻量化通道交互融合模块(CIF),引入SubPixel卷积生成额外双向路径,减少冗余信息。在MS COCO数据集上,相较于Faster-RCNN,大、中、小目标检测精度分别提高了2.2%、1.6%、2.0%。结合MBD模块(多分支膨胀卷积)和MBP模块(多分支池化),保留丰富语义信息,扩大感受野,但不同尺度的特征导致参数量有所增加。

对于模型参数量增加,冷岳峰等^[31]改进Faster R-

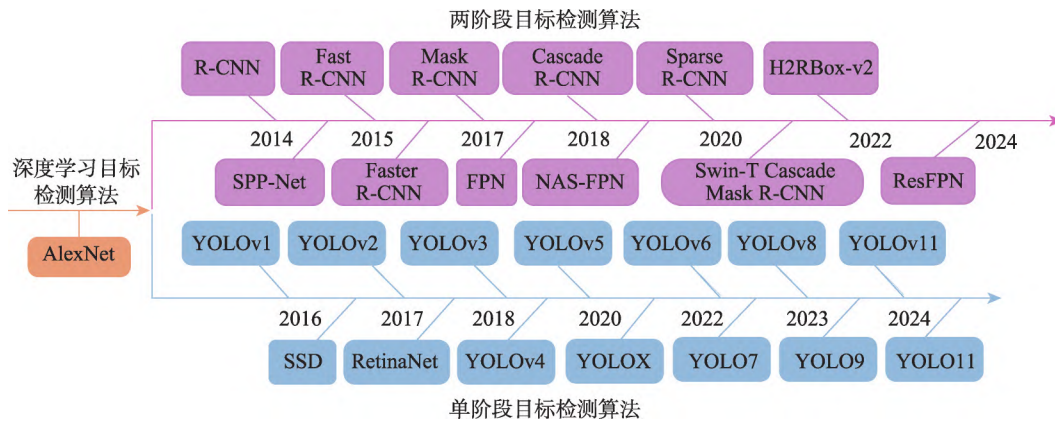


图4 目标检测算法发展历程

Fig.4 Development history of object detection algorithm

表1 两阶段目标检测模型

Table 1 Two-stage object detection model

模型	骨干网络	特点优势	局限性
R-CNN ^[19]	AlexNet	将CNN引入目标检测,实现端到端训练	速度慢,多阶段训练效率低
SPP-Net ^[20]	ZFNet	空间金字塔池化、多尺度特征提取	模型计算复杂度,实时性差
Fast R-CNN ^[21]	AlexNet, VGG16	采用RoI(region of interest)池化、端到端训练,检测速度快	依赖选择性搜索,小目标检测差
Faster R-CNN ^[22]	VGG16, ResNet	引入RPN、共享卷积特征,精度速度提升	模型复杂度高, RPN 计算成本较大
Mask R-CNN ^[23]	VGG16, ResNet101	并行预测掩码与边界框,实例分割精度高	计算复杂度高,推理速度慢
Cascade R-CNN ^[24]	ResNet	多级级联、逐步优化、提高IoU检测精度	推理速度慢,计算成本高
Sparse R-CNN ^[25]	ResNet	引入稀疏注意力机制,无需NMS(non-maximum suppression)后处理,减少冗余计算	动态头复杂,小物体检测弱
Swin-T Cascade Mask R-CNN ^[26]	Swin-Transformer	Swin Transformer结合级联,增强实例分割	难以满足实时性需求,计算资源需求大
FPN ^[27]	ResNet, ResNeXt	特征金字塔融合多尺度信息,小目标检测强	模型结构复杂,依赖骨干网络性能
NAS-FPN ^[28]	ResNet, ResNeXt	自动优化FPN结构,检测分割性能好	搜索过程耗时,需大量计算资源
ResFPN ^[29]	ResNet	残差连接增强特征传递,提升多尺度目标检测精度	计算复杂度提升,特征冗余
H2RBox-v2 ^[30]	ResNet50	对称感知自监督学习, CircumIoU 损失,兼容旋转增强且提高泛化能力	非对称或遮挡物体检测性能下降

CNN,引入特征融合模块、设计轻量化通道注意力模块(lightweight channel attention module, LCAM)将其集成在FPN和RPN之间,减少计算复杂度。改进后的模型在NEU-DET数据集上参数量降低,检测精度和速度分别提升12.6%、40.9%,但实时性仍低于单阶段算法。

因此,研究人员提出了单阶段目标检测算法,以显著提高检测速度,满足实时性需求。

1.2.2 单阶段目标检测

单阶段目标检测采用端到端的预测方式,直接在图像上生成目标的类别和位置信息。主要包括网格划分、锚框预测、多尺度特征融合及联合优化分类与回归损失,典型模型有YOLO、SSD等。这种方法实时性高,结构简单易部署和计算效率优异。但也存在精度相对较低,正负样本失衡导致小目标漏检、依赖锚框设计和复杂场景适应性弱等不足。

YOLOv1首创端到端回归与网格划分。YOLOv2引入锚框、多尺度训练和BN层增强目标适应性。YOLOv3利用Darknet-53和FPN实现多尺度特征融合以改善小目标检测。YOLOv4整合CSPNet、空间金字塔池化(spatial pyramid pooling, SPP)和Mish激活函数并采用Mosaic数据增强提升鲁棒性。YOLOv5针对工程部署优化推出Focus层和轻量化CSP结构。YOLOv6通过RepVGG重参数化和SimOTA动态标签分配平衡速度与精度。YOLOv7以E-ELAN扩展网络和辅助头设计强化特征表达并优化训练效率。YOLOv8借助C2f模块负责使用两个卷积层查找语义分割,平衡任务损失权重。YOLOv9则以可编程梯度信息(programable gradient information, PGI)和广义高效层聚合网络(global efficient layer aggregation network, GELAN)突破深层网络瓶颈,结合动态标签分配及轻量化架构,进一步提升实时性与

复杂场景检测能力。

针对目标检测的实时性,2024年,清华大学团队提出YOLOv10,网络结构如图5所示。它通过无NMS训练,降低延迟实现高效推理。采用整体效率与精度驱动策略,减少模型架构冗余,结合空间-通道解耦下采样技

术,降低计算复杂度。显著提高实时性检测并减少计算成本,但由于架构的选择,对重叠物体检测的准确性相对较低。

对于小目标和重叠目标检测准确性问题,Ultralytics团队提出YOLOv11,如表2所示。使用C3k2(cross stage

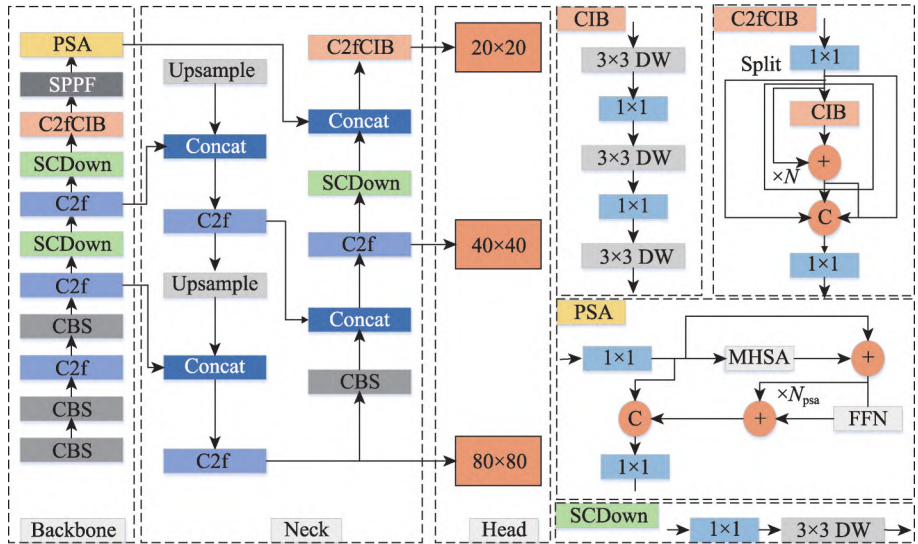


图5 YOLOv10网络结构

Fig.5 YOLOv10 network structure

表2 以经典卷积为架构的单阶段目标检测模型

Table 2 Single-stage object detection model based on classical convolution architecture

模型	骨干网络	特点优势	局限性
YOLOv1 ^[32]	GoogLeNet、VGG-16	端到端、简单高效,首次实现实时检测	小目标检测效果差,定位精度低
YOLOv2 ^[33]	Darknet-19	引入锚框机制、多尺度训练,精度提升	小目标检测性能低,泛化能力不足
YOLOv3 ^[34]	Darknet-53	多尺度FPN、残差网络,平衡速度与精度	模型参数量大,小目标检测仍不足
YOLOv4 ^[35]	CSPDarknet-53	CSPNet结构、Mosaic数据增强,实现实时高精度检测	计算成本高,小目标检测和低功耗设备表现较弱
YOLOv5 ^[36]	CSPDarknet	轻量级设计,推理速度快,部署灵活	复杂场景小目标检测性能弱,精度略低
YOLOX ^[37]	Darknet-53	解耦头、SimOTA动态标签分配,提升精度	小目标检测精度低,对硬件性能要求高
YOLOv6 ^[38]	Efficient Rep	轻量化设计、自蒸馏策略,检测速度快	网络结构复杂度高,需精细调参
YOLOv7 ^[39]	Extended-ELAN	扩展策略、重参化、标签分配,高效训练	内存占用大,复杂场景需平衡实时与精度
YOLOv8 ^[40]	Darknet-53	改进损失函数、动态锚框,精度和速度高	小目标检测弱,硬件要求高
YOLOv9 ^[41]	GELAN	可编程梯度提升可靠性,轻量架构高效	依赖特定数据集,训练复杂度高
YOLOv10 ^[42]	Darknet-53	无NMS训练、解耦下采样,减少冗余计算	模型复杂度高,训练成本高
YOLOv11 ^[43]	C3k2/C2PSA	高效特征融合、跨平台部署	硬件要求高,小目标检测精度低
SSD ^[44]	VGG16	多尺度锚框、端到端训练	小目标检测效果差
DSSD ^[45]	ResNet-101	深层特征增强,提升小目标检测精度	计算量增加,速度下降
RSSD ^[46]	VGG-16	优化残差结构,减少梯度消失	模型复杂度高,推理速度慢
FSSD ^[47]	VGG-16	轻量级特征融合模块,增强多尺度表达	特征语义冲突,部署兼容性差
LMUAV-YOLOv8 ^[48]	AD-FTA-Darknet	轻量化多尺度融合,高效、实时性强	计算量增加,小目标检测精度低
YOLOX改进 ^[49]	CA-CSPDarknet	轻量化设计,多方向检测精度高,实时性强	复杂背景及恶劣天气鲁棒性差
RetinaNet ^[50]	ResNet-50、ResNet-101	Focal Loss解决样本不平衡	训练复杂,推理速度慢
CenterNet ^[51]	Hourglass Network	关键点检测、无NMS,速度快	依赖中心关键点,复杂场景受限
FCOS ^[52]	ResNet	无锚框、Center-ness修正,减少超参数	边界定位精度不足,小目标检测效果差
SqueezeDet ^[53]	SqueezeNet	轻量化设计、实时强,适合嵌入式设备	复杂场景检测精度低
EfficientDet ^[54]	EfficientNet、BiFPN	高效特征融合、复合缩放,精度高	依赖预训练,资源约束场景优化不足

partial with kernel size 2)模块,提高了计算效率和推理速度。引入C2PSA(convolutional block with parallel spatial attention)模块优化特征提取。显著改善小目标和重叠目标检测的准确性,但也增加了实现复杂度、更高的训练成本和资源需求。

为解决上述问题,基于YOLO系列的轻量化算法相继提出。如董一兵等^[48]提出LMUAV-YOLOv8算法。引入多尺度轻量化特征融合网络(UAV_RepGFPN),提升特征融合效果,在VisDrone2019数据集上参数量减少31.5%。设计上下文指导的可编程梯度信息策略(UAV_PGI)和轻量化特征提取模块(FTA_C2f),显著提升空间特征捕捉能力和特征提取效率,相较于原模型mAP提升5.1%。但由于增加小目标检测层,模型计算量增加。雷景生等^[49]对YOLOX进行改进。引入通道注意力机制(coordinate attention, CA),增强特征提取能力并减小噪声干扰,在CCPD数据集上参数量减少27.2%。使用轻量级路径聚合网络(LWPAN)优化计算效率,相较于原模型计算量减少9.1%,mAP提升3.0%。此外,RetinaNet、CenterNet和FCOS等模型目标检测性能也比较优秀。

除上述以经典卷积为架构的算法之外,基于Transformer为架构的DETR系列目标检测模型,具有高精度和鲁棒性,但训练时间长、推理速度较慢和计算资源需求高。

DETR是首个基于Transformer的端到端目标检测模型,由于Transformer对全局特征提取的突出优势,研究者对不同角度的DETR进行改进,如表3所示。

为解决小目标检测较差和模型收敛过慢,采用改进注意力机制,提出Deformable DETR和Dynamic DETR算法;通过解码器特征信息之间的交互改进,提出Conditional DETR和DAB-DETR算法,优化了查询和检测的精度,在此基础上针对标签匹配方式,提出DN-DETR提高了模型收敛速度以及检测精度。

Liu等^[69]为进一步优化小目标检测、训练收敛速度慢的问题,提出F-DETR。通过引入异构多尺度特征和多分支结构,实现局部与全局特征的交互,加速训练收敛,减少计算冗余。在COCO val2017数据集上训练速度提升14倍,小目标检测精度提升6.2%。但多尺度特征的结合增加计算资源需求,使实时检测场景的部署受限。

表3 以Transformer为架构的单阶段目标检测模型

Table 3 Single-stage object detection model based on Transformer architecture

模型	相关改进	优势	局限性
DETR ^[55]	采用二分匹配损失和Transformer架构进行直接预测	架构简单易实现,显著提升大目标检测性能	训练复杂耗时长,小目标检测效果差且检测数量有限
Deformable DETR ^[56]	引入可变注意力机制和可变形卷积,迭代边界框优化	减少冗余计算,收敛快、性能好、效率高且多尺度融合能力强	由于采用了多尺度,模型结构复杂,计算成本高
Dynamic DETR ^[57]	引入动态注意力机制,卷积动态编码器和基于RoI的动态解码器	显著提升小目标检测性能且训练收敛速度快,同时提高了检测精度	计算复杂度高、极端场景适应性差及数据依赖性强
Conditional DETR ^[58]	提出条件交叉注意力机制,引入条件空间查询	提高收敛速度,降低了对内容嵌入的依赖性和训练难度	小目标检测弱、复杂场景鲁棒性不足,依赖高算力资源
DAB-DETR ^[59]	采用动态锚框作查询、宽度与高度调制的高斯位置注意机制	加快训练收敛速度,增强不同尺度目标检测性能	小目标检测精度较差,计算复杂度高,泛化能力受限
DN-DETR ^[60]	引入查询去噪任务及架构调整加速训练	加速模型收敛、提升模型精度高,通用性强	超参数敏感、计算资源高、动态适应性不足
SD-DETR ^[61]	通过解耦特征、生成任务感知查询和调整损失函数改进	性能提升显著、适应性广且改善分类和定位任务的特征与预测不对齐	增加了模型复杂度,计算资源与部署成本高
DA-DETR ^[62]	融合CNN与Transformer特征,设计CTBlender实现信息融合	提升无监督域自适应目标检测的性能且泛化能力强	二分图匹配的不稳定性导致资源消耗高、实时性受限
Lite-DETR ^[63]	设计交错更新编码器块并引入关键感知变形注意力(key-aware deformable attention, KDA)	优化特征处理,降低计算成本,同时维持高检测性能且泛化性好	实时推理不足,存在跨模态特征融合受限
Focus-DETR ^[64]	设计基于多尺度特征的token评分机制,双注意力机制重构编码器	提高检测精度,减少模型复杂度,降低计算成本,通用性强	效率与精度失衡、级联误差累积,局部注意力受限且小目标检测性能差
RT-DETR ^[65]	设计高效混合编码器,提出不确定性最小化查询选择	高精度、高处理速度、实时检测,无NMS且能灵活适应场景	计算资源需求高,复杂场景泛化能力弱,影响实时部署
MS-DETR ^[66]	采用混合监督并调整解码器结构及预测模块	提升实例分割性能,计算内存高效且收敛速度快	混合监督增加训练时间和计算复杂度,需精细调参
Mobile-DETR ^[67]	用MobileNetv3替换特征网络并引入CBAM模块	显著提升了检测精度、模型轻量化和训练效率	在检测精度方面有待提高
DR-DETR ^[68]	设计CSDID机制、FGAF-Fusion模块,采用DCNv3可变形卷积	提高小目标检测能力,计算复杂度和参数量低,鲁棒性和泛化能力强	模型复杂度和检测速度方面仍有不足

为应对高计算成本,Chen等^[70]提出DHS-DETR模型。设计一种动态头部切换(dynamic head switching, DHS)策略,能够根据图像难度自适应地选择不同复杂度的Transformer头。在MS COCO数据集上计算量减少25.9%。引入在线头部蒸馏(online head distillation, OHD),减少因动态切换带来的性能损失,轻量级头的精度提升7.0%。

此外,基于DETR改进的Lite-DETR、RT-DETR(real-time DETR)和Focus-DETR等模型,说明其在目标检测算法轻量化领域发展空间很大。

为进一步减少模型参数量并加快训练速度,王鹏九等^[67]改进Deformable DETR模型。采用轻量级特征提取网络MobileNetv3,显著减少模型参数,提高训练速度。在ABOships数据集上参数量减少66.5%,检测速度提升92.8%。结合CBAM(convolutional block attention module)注意力机制,增强特征提取能力,相较于原模型mAP提升13.4%,但改进后的算法精度仍有提升空间。冯永安等^[68]提出新型DR-DETR模型,其网络结构如图6所示。采用通道和空间双重细化的信息蒸馏机制(channel-spatial dual refinement information distillation mechanism, CSDID),减少网络对冗余特征的处理。设计了特征信息门控自适应融合模块(FGAF-Fusion),提高模型对小目标的检测能力,在RDD2022数据集上相较于RT-DETR模型mAP提升4.2%。

1.3 总结分析

传统目标检测依赖手工特征和多阶段处理且存在计算效率低、泛化能力弱等问题,难以应对复杂场景。深度学习方法通过端到端架构实现了突破。两阶段算法检测精度高,但速度较慢;单阶段算法则通过简化流程显著提升实时性,但小目标检测能力不足;而基于Transformer模型(DETR系列)通过全局特征提取提高精度,但仍需解决训练耗时和计算成本问题。

针对传统目标检测的效率与泛化瓶颈,结合深度学习发展趋势,当前研究目标检测算法的轻量化,将成为平衡实时性与精度的关键技术路径。

2 轻量化评价指标

对于目标检测算法轻量化评估,关键在于确定其准确性和速度,以及与其他模型比较。为验证算法检测效果,采用参数量、计算量、每秒帧数(frame per second, FPS)、平均精度(mean average precision, mAP)等作为评估算法轻量化各项性能的指标。

2.1 模型效率与复杂度指标

(1) 参数量

参数量(parameters)是模型中可训练参数的总数,直接影响模型的存储需求和内存占用。

(2) 计算量

计算量(giga floating point operations, GFLOPs)是模型在进行一次前向传播时的浮点运算数量,直接影响

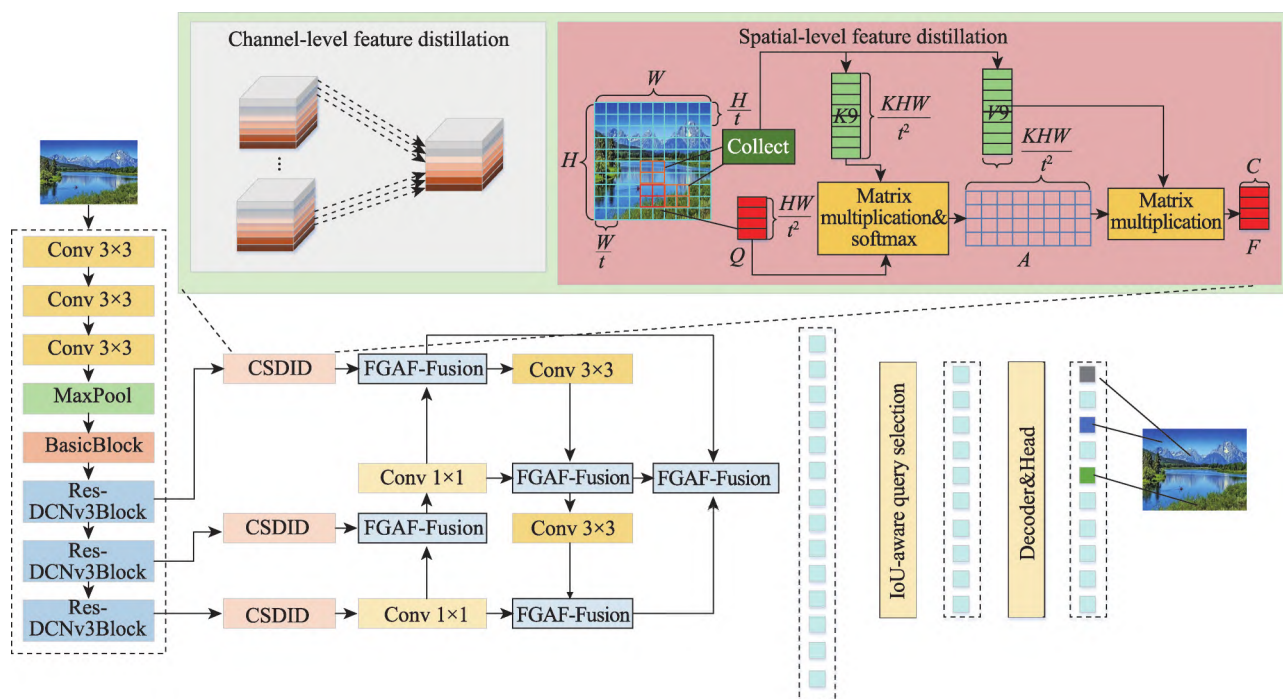


图6 DR-DETR网络结构

Fig.6 DR-DETR network structure

设备能耗和推理速度。

(3) 帧率

帧率FPS表示模型每秒钟能够处理的图片数量。帧率值越高模型检测速度越快。计算方式如式(1),其中 t 为单张图片所需的处理时间。

$$FPS = \frac{1}{t} \quad (1)$$

2.2 检测精度指标

(1) 准确率和召回率

准确率(Precision)衡量模型预测结果可靠性(式(2))。高准确率意味着模型预测为正样本比例较高。召回率(Recall)衡量模型检测缺陷的完整性(式(3))。高召回率意味着模型能够检测出大多数实际存在的缺陷。

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

式中, TP 为正样本, FP 为负样本, FN 指的是模型错误地将一个正样本检测为负样本。

(2) 平均精度和平均精度均值

平均精度(average precision, AP)表示单类别精确度,数值越大说明精度越高。计算方式如式(4),其中 R 表示不同交并比下的召回率。

$$AP = \int_0^1 P(R) dR \quad (4)$$

平均精度均值mAP表示多类别精确度。计算方式如式(5),其中 K 为类别总数。

$$mAP = \frac{1}{K} \sum_{i=1}^K AP_i \quad (5)$$

2.3 评价指标应用分析

2.3.1 嵌入式设备

嵌入式设备需要低功耗、实时性以及模型体积小。其中各项评价指标需求优先级为:FPS方面需满足实时处理($FPS \geq 30$),如无人机避障就需要高帧率;参数量与模型大小要适配ARM CPU或低端GPU,如YOLOv3-Tiny在CPU上仅6.2 FPS,需压缩至MB级;GFLOPs要降低计算能耗,如MobileNet-SSDv2通过深度可分离卷积减少GFLOPs;mAP则是在保证实时性的前提下追求精度,如农业检测中 $mAP \geq 90\%$ 。

2.3.2 工业检测

工业检测需要高精度、稳定性、多目标处理。其中各项评价指标需求优先级为: $mAP@0.5:0.95$ 要达到严格定位精度,防止如电子元件检测中因 $IoU < 0.8$ 而出现微小缺陷漏检的情况;Recall需减少漏检率,像钢铁缺陷检测场景中就需覆盖所有潜在目标;FPS与预处理时间需平衡速度与精度,如工业流水线可接受的帧率为10~20 FPS。

2.3.3 自动驾驶

自动驾驶需要实时性、多尺度目标检测、复杂场景适应性。其中各项评价指标需求优先级为: $mAP@0.5$ 要求能快速检测车辆和行人, $IoU=0.5$ 的情况可以执行;FPS方面需达到30($FPS \geq 30$)以支持实时决策,如YOLOv4在GPU上可达65 FPS;小目标检测能力通过AP来评估,如卫星图像中车辆检测及交通标志识别。

3 目标检测算法轻量化途径

虽然CNN、YOLO、SSD系列算法在目标检测方面效果很好,但其普遍具有较高计算复杂度和较大模型体积,无法满足计算能力、存储空间、功耗等资源受限场合的应用需求。为提高目标检测算法在资源受限条件下的性能,各种轻量化目标检测算法先后发展而来,如轻量级网络MobileNet、ShuffleNet、FasterNet等,轻量级卷积设计或使用网络剪枝、知识蒸馏、神经网络搜索(neural architecture search, NAS)等技术压缩模型实现网络轻量化部署,有效提高了目标检测算法的性能水平。

3.1 轻量级网络

随着网络层数的加深,模型的性能变得越来越好,随之而来的问题是模型参数量增加和推理速度减缓,因此轻量化网络的需求愈加强烈。轻量化网络的设计核心是在尽可能保证模型精度的前提下,降低模型的计算复杂性和空间复杂度。

3.1.1 MobileNet网络结构

Google团队2017年推出了MobileNet系列轻量级卷积神经网络架构。MobileNet^[71]是专为移动和嵌入式设备设计。采用深度可分离卷积(depthwise separable convolution, DSC),降低计算复杂度。引入两个全局超参数,从而进一步提高速度并减小模型尺寸。但该模型网络结构简单导致特征学习能力受限。

2018年,Sandler等^[72]提出MobileNetv2。通过引入倒残差结构(inverted residual)和线性瓶颈(linear bottleneck)层来提高网络表征能力,结合DSC显著减少计算量并提高精度,ResNet残差结构、MobileNetv2倒残差结构如图7所示。但线性瓶颈限制复杂特征的表达能力。

MobileNetv3^[73]与MnasNet^[74]创建所用方法相同,但有所修改。通过结合硬件感知神经架构搜索(NAS)和NetAdapt算法,优化网络结构。引入SE(squeeze-and-excitation)自注意力模块和HS(h-swish)激活函数提高模型精度。使用NAS技术设计MobileNetv3-large模型和MobileNetv3-small模型分别部署高资源、低资源用例。但SE模块的全局平均池化导致信息丢失,特征提取能力受限制。

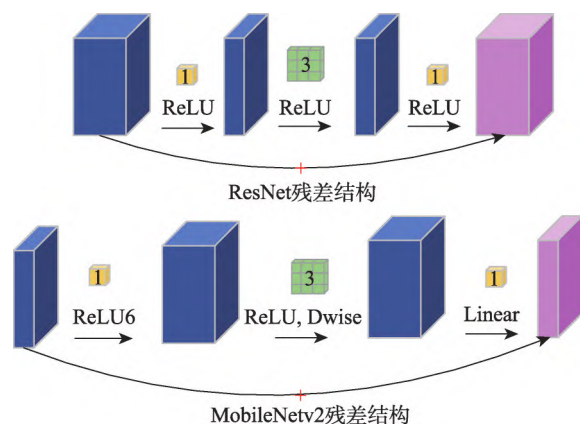


图7 残差模块结构

Fig.7 Residual module structure

2024年, Google团队提出 MobileNetV4^[75]。该网络引入通用倒置瓶颈UIB(universal inverted bottleneck)搜索块和移动版多头注意力(mobile MQA)两个创新模块, 统一不同微架构, 共享所有头的键和值, 提高了计算效率。并结合两阶段NAS提高搜索效率, 显著提升模型准确性。

针对模型复杂度高和参数量多, 肖振久等^[76]提出轻量化网络架构 sep-res18-s3。在特征提取模块上结合 DSC 和 ResNet 残差结构, 融合不同尺度的信息且防止梯度消失。相比设计的类残差网络 cov-res18-s3, 该网络仍能降低 58.33% 的参数量和 81.82% 的计算量。Junayed 等^[77]通过新的特征融合模块, 结合深度卷积骨干和多尺度注意力机制, 显著提升模型在外观复杂度变化时的处理能力。但是 1×1 逐点卷积增加了内存消耗, 此外逐点计算难以实现并行加速。

总体看来, 在 MobileNet 系列网络中都大量使用了深度可分离卷积作为基础卷积模块, 这种方法显著降低卷积运算次数, 提高神经网络训练和推理效率。

3.1.2 ShuffleNet 网络结构

2017年, 旷视科技提出 ShuffleNet^[78], 网络结构如图8所示。其创新在于引入逐点组卷积(pointwise group convolution), 将所有 1×1 卷积替换为 1×1 组卷积, 结合通道混洗(channel shuffle)操作, 在保持模型精度的同时显著降低计算成本。通道混洗输入输出通道数不同, 组数过多导致内存消耗, 实际运行速度受限。

2018年, Ma 等^[79]提出 ShuffleNetV2, 其建立在四条轻量化准则之上。通过引入通道分裂(channel split)进行深度卷积和逐点组卷积, 然后采用通道混洗使分组后的特征得到融合和交流。使模型在保持精度的同时, 显著提高网络效率。

为优化检测精度和效率, Yang 等^[80]提出轻量化的

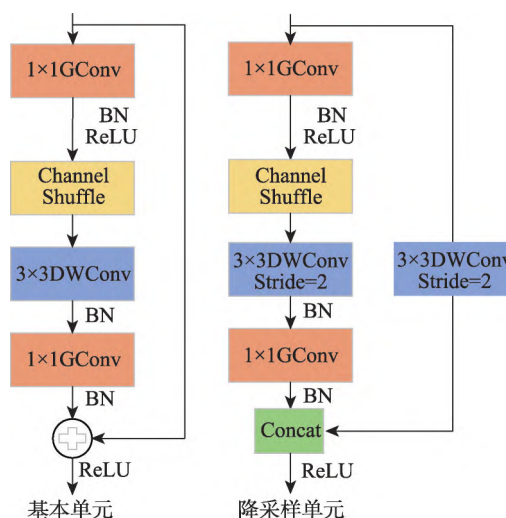


图8 ShuffleNetV1 网络结构

Fig.8 ShuffleNetV1 network structure

PSE-ShuffleNetV2。采用 SE 模块和重新选择 PReLU 激活函数, 显著提升检测准确性和实时性能。相较于原模型 mAP 提高 2.78%, 实时检测准确率提高 3.75%。

为实现模型低复杂度和高精度, Zhang 等^[81]在模型中加入 MBConv 结构和 SE 模块, 实现特征融合, 过滤复杂的信道特征, 降低模型复杂度。改进后的 ShuffleNetV2 模型在 CCSR NIR、CASIA NIR-VIS 2.0 F 和 Oulu-CASIA 数据集上, 准确率分别达到 98.50%、99.87% 和 100%。引入自适应 Gamma 矫正技术提高泛化能力。

3.1.3 GhostNet 网络结构

2016年, Iandola 等^[82]提出一种更小、更智能的网络 SqueezeNet。该架构通过用 1×1 滤波器替换 3×3 滤波器, 将模型参数量降至 AlexNet 的 1/50, 同时保持了基准精度。随后 MobileNet 和 ShuffleNet 系列轻量级网络也取得一定成果, 但仍存在一些不足。

针对上述问题, 华为诺亚方舟实验室在 2019 年提出 GhostNet^[83]。它通过引入 Ghost 模块和 Ghost 瓶颈结构, 实现高效特征提取能力, 其结构如图9所示。GhostNet 在保持高精度的同时降低计算成本。但受限于网络深度和空间复杂度, 对复杂任务的适应性较弱。

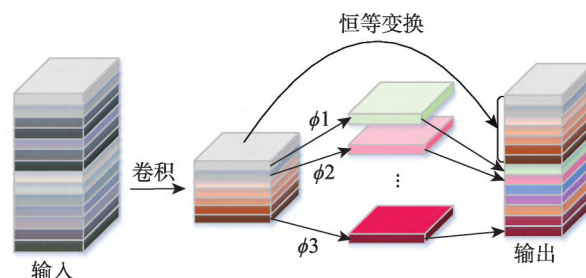


图9 Ghost 模块结构

Fig.9 Ghost module structure

2022年,在GhostNet基础上提出GhostNetv2^[84]。该模型采用改进的Ghost模块,提出硬件友好型注意力机制(DFC),将完全连接(fully connected, FC)层分解为水平和垂直FC层。提升在轻量级卷积神经网络中捕获长距离空间信息能力,适用于在资源受限的移动端,但注意力图进行缩放的操作影响性能稳定性。

GhostNetv3^[85]是在GhostNet系列上的重要升级版本。将轻量级模型的重参数化与知识蒸馏结合,提高模型性能。创新动态机制,提出了适合轻量级模型的学习率调整策略,在保持推理延迟不变的同时显著提升模型精度。在轻量化与性能之间实现了显著突破。

为提高模型分割性能,Ardiyanto等^[86]引入ESD-NET。采用GhostNet融合块GFB(ghostNet fusion block)减少下采样空间信息丢失,提升分割精度。结合GAL(global auxiliary layer)提高特征和信息融合能力。优化后的ESDNet模型在SD-saliency-900、Magnetic-tile公共表面缺陷数据集上,mIoU(mean intersection over union)分别提高9.5%和30%。尽管增加少量网络参数和计算延迟,但这些改进使得模型在嵌入式设备上也能有效运行。

针对计算资源和实时性检测不足,王迎龙等^[87]提出BG-YOLO算法。使用GhostNetv2替换传统卷积模块,显著减少模型参数量。在Det-Fly数据集上相比原模型参数量减少33.2%。引入双层路由注意力机制(BiFormer)捕捉长距离依赖关系,相较于原模型检测精度和帧率分别提升10.3%和2.4%。显著提升了检测精度和实时性。

3.1.4 EfficientNet网络结构

2019年,Google AI研究团队提出EfficientNet^[88]。引入基于固定扩展系数的复合扩展方法,利用神经架构搜索(NAS)设计了新的基准模型,并结合复合扩展方法创建了EfficientNet-B0~B7模型家族。显著提升了模型在ImageNet及其他迁移学习数据集上的准确性与效率,但模型复杂度也在逐步增加,且大尺寸图像训练速度缓慢,NAS成本高。

为了进一步优化训练速度与泛化能力,2021年推出了EfficientNetv2^[89]。通过结合训练感知的NAS和模型扩展,并引入融合运算(Fused-MBConv)和改进的渐进式学习策略。在ImageNet和CIFAR等数据集上实现了更快的训练速度以及更好的迁移学习表现,但Fused-MBConv在部分硬件加速器上可能效率不足。

卜晓燕等^[90]通过在EfficientNet主干网络中引入注意力机制(CBAM),使模型更精细地提取特征。优化双向特征金字塔网络(BiFPN)的层次结构和特征融合策略,提升多尺度特征提取能力。在自建数据集上的检测

精度达到91.32%,相较于EfficientDet-D0和YOLOv5s分别提升了3.69%和2.05%,但计算量有所增加,检测速度略有下降。

对于上述问题,田栋等^[91]采用EfficientNetV2替代原有特征提取模块,并在检测头中引入BiFPN特征融合模块,减小网络结构,提升特征提取能力。在KITTI数据集上参数量和计算量分别降低39%和40%。使用Meta-ACON激活函数优化模型收敛性,训练更高效,相较于原模型检测精度提高了1.6%。

3.1.5 FasterNet网络结构

针对冗余计算和内存访问,2023年Chen等^[92]提出新型神经网络FasterNet。该网络使用一种新的卷积模块-部分卷积(partial convolution, PConv)解决传统内存访问频繁问题,同时构建以PConv和PWConv为基础的简单架构。其结构包含四个层次阶段,各阶段有FasterNet块,块内PConv层后接两个PWConv层,采用合理的归一化和激活层设置。同时为了在不同的计算资源上广泛应用,提供了多个版本,分别为FasterNetT0/T1/T2、FasterNet-S/M/L。它们具有相似的架构,但深度和宽度有所不同。然而,过度简化归一化和激活层会限制特征表达能力,需要权衡性能与多样性。

为平衡模型性能,Guo等^[93]通过FasterNet模块替代YOLOv8s的主干网络,减少参数量。在URPC200数据集上模型体积减少24%。改进双向特征金字塔网络,并利用GSConv和部分卷积模块替换标准卷积和瓶颈模块,简化特征层优化融合方式,提升检测性能。相较于原模型检测精度提升1.2%,但面对复杂环境适应性不足。

针对复杂环境速度和精度的平衡,Dong等^[94]提出FCCD-SAR算法。引入FasterNet作为主干网络结合轻量化上采样算子CARAFE,有效平衡模型的精度和计算效率。在MSTAR数据集上参数量和计算量分别减少14.9%和23.8%。采用DyHead注意力机制检测头,提升表达能力,相较于原模型检测精度提升1.4%。

3.1.6 StarNet网络结构

2024年,微软公司提出StarNet^[95]轻量化网络,其网络结构如图10所示。它采用传统的4级分层网络结构,利用卷积层进行下采样,并通过Star Blocks进行特征提取。深度卷积层之后为批量归一化取代层归一化。StarNet的关键在于其星运算(star operation),具有将输入映射到高维的非线性特征空间的能力。该方法建立的StarNet网络参数量少,可实现较高的检测精度,但调整灵活性不如传统网络,可能会限制模型的进一步优化。

为进一步优化,Chu等^[96]提出YOLO-SDS算法。通过引入StarNet模块替代YOLOv8的骨干网络,实现高维非线性空间映射,在NEU-DET数据集上参数量减少34.4%。颈部网络结合轻量化模块DWR与特征提取模

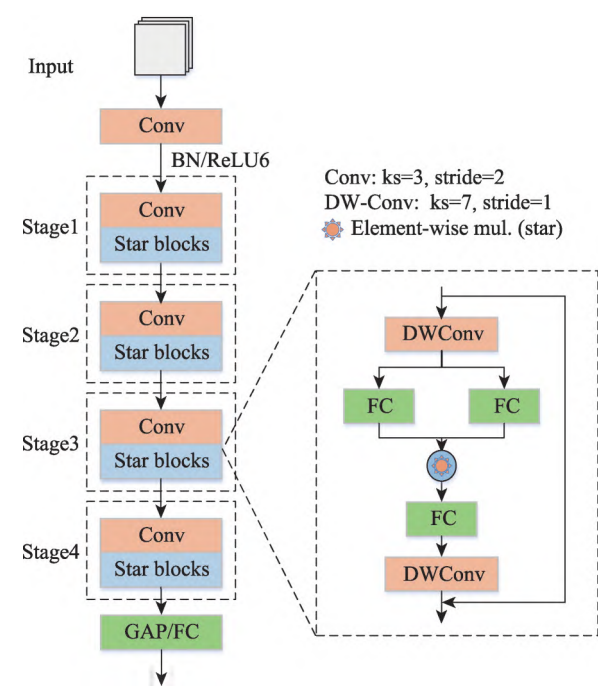


图 10 StarNet 网络结构

Fig.10 StarNet network structure

块 C2f, 提升多尺度特征提取效率, 改善复杂环境检测性能, 相较于原模型检测精度提升 1.5%, 但在实时性缺陷检测上, 性能有待提升。

针对实时性检测参数量大和计算复杂, 古莹奎等^[97]提出 SGHS-DETR。采用 StarNet 主干网络降低模型复杂度, 在自建数据集上参数量降低 65.5%。结合改进的 HWD 模块与 ShapeIoU 损失函数, 加快推理速度同时提高回归精度, 相较于原模型计算量降低 72.1%, 检测精度提升 1.5%。

3.1.7 Neck 网络结构

2019 年, Google AI 团队在 EfficientDet 目标检测算法中引入了双向特征金字塔网络 (BiFPN)^[54]。通过构建自顶向下和自底向上的双向信息流通道, 动态调整不同尺度特征之间的权重, 实现多尺度特征的反复融合, 提高多尺度的检测能力。但堆叠多个 BiFPN 模块或使用 softmax 加权可能增加计算量, 影响 GPU 效率。本文选择仅在颈部网络进行优化, 构建轻量化颈部网络。

葛雯等^[98]提出 MGL-YOLO 算法。通过在颈部网络中引入 GLSA (global-local spatial attention) 和 GSConv 模块改进 BiFPN, 增强特征处理能力, 同时实现轻量化。在 VEDAI 数据集上参数量和计算量分别降低 47% 和 32%。沈骞等^[99]结合 FasterBlock 和 CGLU 模块改进颈部网络, 减少计算量优化提取特征, 实现颈部网络的轻量化。

同样对于颈部网络的轻量化, 罗显志等^[100]采用 EMSFPN (efficient multi-scale FPN) 结构改进颈部网络, 降低参数量。同时结合 inner-WIoU 损失函数优化边框回归,

提高模型在复杂定位中的精度。模型参数量减少 9.8%, 检测精度提升 9.1%, 确保模型在有限资源下的高效运行。

3.1.8 Head 网络结构

针对头部网络的轻量化, 沈骞等^[99]在检测头中加入轻量级共享卷积检测头 (lightweight shared convolutional detection head, LSCD), 有效减少参数量, 降低模型复杂度, 提高了计算性能。在 StateFarm 数据集上参数量和计算量分别减少 23.3% 和 26.8%。实现高效检测, 同时在嵌入式平台上表现优异, 葛雯等^[98]采用轻量化的 Light-head 检测头优化检测头结构, 同样实现头部网络的轻量化。

不仅如此, 史丽晨等^[101]提出 CDD-YOLO。引入动态头部框架 (dynamic head) 和剪枝算法 GroupNorm, 减少复杂背景与冗余参数实现轻量化。在自建数据集上参数量和计算量分别减少 83.8% 和 86.6%。替换边界框损失函数 DPlIoU, 提高精度和泛化能力, 相较于原模型检测精度提升 3.6%。但动态计算可能降低推理速度, 过度剪枝导致精度损失。

对于上述问题, 钱尚乐等^[102]采用 MobileNetv3 主干网络, 在自建数据集上参数量和计算量分别减少 11.7% 和 21.2%。结合 WIoU 损失函数并在检测头中嵌入 GS 解耦头 (gsconv decoupled head, GSDH), 提升收敛速度和并行计算速率, 相较于原模型检测精度提升 10.1%, 适合轻量化部署。表 4 总结了不同轻量级网络的性能对比。

表 4 轻量级网络性能对比

Table 4 Lightweight network performance comparison

模型	参数量/ MB	计算量/ GFLOPs	精度 (Top-1)/%	FPS
MobileNetv1 ^[71]	4.2	569	70.6	150
MobileNetv2 ^[72]	3.4	300	72.0	200
MobileNetv3 ^[73]	2.5	40	67.5	350
MobileNetv4 ^[75]	11.7	5.6	87.0	263
ShuffleNetv1 ^[78]	0.5	140	78.5	300
ShuffleNetv2 ^[79]	2.2	140	69.7	300
SqueezeNet ^[82]	1.2	83	57.5	450
GhostNetv1 ^[83]	5.2	141	73.9	280
GhostNetv2 ^[84]	6.1	167	75.3	260
GhostNetv3 ^[85]	6.1	169	77.1	70.3
EfficientNetv1 ^[88]	5.3	0.39	76.3	220
EfficientNetv2 ^[89]	24.0	8.8	83.6	80
FasterNet ^[92]	7.6	0.85	76.2	230
StarNet ^[95]	2.9	425	73.5	270
BiFPN ^[98]	0.88	0.67	43.9	60.8
G-YOLOv8n ^[99]	1.6	4.8	99.4	110
EMSFPN ^[100]	5.8	11.8	72.5	75.4
CDD-YOLO ^[101]	4.2	10.6	89.6	257
G-YOLOv7-tiny ^[102]	5.3	10.4	96.6	68.8

3.1.9 总结分析

综上所述,轻量级网络包含对骨干网络、颈部网络、头部网络的设计优化。轻量级骨干网络通过深度可分离卷积、倒置残差、星运算等创新技术,在降低计算复杂度与模型尺寸的同时保持了较高精度。如FasterNet提出PConv和PWConv组合架构,通过减少内存访问提升推理速度。StarNet采用star operation构建分层网络,在

低参数量下实现较高检测精度。轻量级颈部和头部网络的优化是实现端侧高效推理的重要补充。颈部网络如BiFPN通过双向特征金字塔与动态权重分配提升多尺度检测性能,采用通道注意力替代全局池化。头部网络可通过精简检测分支、引入轻量级激活函数,HSwish替代ReLU及优化损失函数GIoU、CIoU减少计算成本。表5总结了轻量级网络架构设计特点。

表5 轻量级网络架构设计
Table 5 Lightweight network architecture design

模型	机制	优势	局限性	应用场景
MobileNetv1 ^[71]	采用深度可分离卷积结合宽度和分辨率乘数	显著减少参数和计算成本,适配不同场景,支持多任务部署	结构简单导致特征学习弱,激活函数易丢失低维信息	自动驾驶、工业检测
MobileNetv2 ^[72]	引入倒置残差与线性瓶颈,结合深度可分离卷积	高效轻量,平衡精度与速度,移动端实时推理和多任务部署	低维特征易丢失,复杂任务精度受限	自动驾驶、工业检测、医学影像
MobileNetv3 ^[73]	结合NAS与结构优化,引入轻量模块和高效激活函数	高精度低延迟,适配移动端硬件,计算效率高	小数据易过拟合,部分硬件部署性能受限	自动驾驶、无人机检测、医学影像
MobileNetv4 ^[75]	整合UIB、Mobile MQA与两阶段NAS	跨硬件兼容性强,高精度,延迟低且无需平台适配	复杂结构可能降低部署效率	自动驾驶、工业检测
ShuffleNetv1 ^[78]	引入逐点组卷积和通道混洗	计算高效,参数量少,移动端推理快,特征交互增强模型表达	重排操作效率低,复杂任务性能受限,超参数敏感,扩展性有限	行人检测、工业检测、医学影像
ShuffleNetv2 ^[79]	遵循四项轻量化准则,引入通道分裂结合通道混洗,特征重用	计算效率高,平衡速度与精度,适配移动端轻量化部署	在小目标检测、复杂背景处理中的性能较弱	自动驾驶、工业检测、无人机检测、医学影像
SqueezeNet ^[82]	Fire模块结合网络剪枝、量化和延迟下采样	参数量少且计算高效,适配移动端与边缘设备部署	精度受限,并行效率低,依赖后处理且训练稳定性不足	自动驾驶、工业检测、医学影像
GhostNetv1 ^[83]	线性变换生成冗余特征,结合残差结构减少传统卷积计算量	计算高效,参数量较少,适配移动端且模块通用性强	内存占用较高,复杂任务泛化性弱且依赖精细调参	自动驾驶、工业检测、无人机检测、医学影像
GhostNetv2 ^[84]	引入DFC注意力机制,结合硬件友好设计优化特征聚合	全局特征高效融合,计算量低,适配移动端轻量化部署	线性操作限制复杂特征建模,下采样损失细节	交通检测、工业检测、农业检测
GhostNetv3 ^[85]	采用重新参数化、知识蒸馏与优化训练策略	高效计算与硬件适配性强,在低资源下实现高精度低延迟推理	依赖教师模型性能,需调整超参数,轻量化设计模型性能受限	自动驾驶、工业检测、无人机检测
EfficientNetv1 ^[88]	复合缩放深度/宽度/分辨率,结合MBConv模块与NAS	高效平衡性能与资源,参数少速度快,泛化能力强	依赖初始模型设计,浅层计算效率低且NAS成本高	小目标检测、医学影像
EfficientNetv2 ^[89]	融合Fused-MBConv和渐进式学习,优化非均匀缩放	训练速度快,轻量高效,硬件适配和泛化能力强	依赖训练数据规模,超大图像处理受限,部分硬件效率不足	工业检测、医学影像
FasterNet ^[92]	PConv减少冗余,结合倒残差和层级优化结构	跨平台,高精度,推理速度显著提升,结构简单适配多任务	特征多样性受限,复杂场景敏感,小数据易过拟合	工业检测、交通检测、农业检测
StarNet ^[95]	通过星运算映射高维特征,结合分层架构提升效率	高效性能、低延迟且无需激活函数,适配移动端部署	结构简单限制高阶任务优化,扩展性和硬件适配能力较弱	工业检测、自动驾驶
BiFPN ^[98]	双向路径加权融合多尺度特征,优化结构提升效率	高效融合特征,参数少速度快,适配多资源场景且显著提升检测精度	设计复杂,依赖复合缩放,硬件适配性受限	小目标检测、自动驾驶、工业检测、医学影像
G-YOLOv8n ^[99]	引入共享卷积(LSCD)减少参数冗余	计算高效,参数量少,泛化能力强	复杂场景性能下降,小目标及遮挡识别能力不足	自动驾驶
EMSFPN ^[100]	引入双向特征融合与多尺度模块,增强特征表达能力	高效融合特征,提升检测精度与效率,适合资源受限场景	计算量与复杂度增加,实时性差	小目标检测、无人机检测
CDD-YOLO ^[101]	融合CDConv、坐标注意力、DyHead和剪枝算法	轻量高效,收敛速度快,增强多尺度和视角变化的适应性	计算成本高,过度剪枝降低精度,超参数调整影响稳定性	工业检测、自动驾驶
G-YOLOv7-tiny ^[102]	采用轻量化网络、注意力机制、优化损失函数和解耦头	增强目标区域特征学习和并行计算效率,模型鲁棒性强	单独使用检测精度受限,泛化能力不足	工业检测、小目标检测

3.2 轻量级卷积设计优化

随着网络性能和复杂度的提高,网络中会出现大量冗余。因此设计出更合理、高效的网络结构对于提升模型性能具有重要意义。如通过调整卷积核大小、改进卷积结构等方面来构建轻量级卷积神经网络。

3.2.1 调整卷积核大小

在早期的卷积神经网络中,AlexNet^[103]采用 11×11 和 5×5 等大尺寸卷积核提升图像分类性能。但后续研究发现其存在明显局限性,大卷积核显著增加模型计算量和训练难度,易丢失局部细节特征,且限制网络结构灵活性。针对这一问题,研究者逐步转向更高效的卷积核优化策略,采用小卷积替换大卷积和使用特殊卷积“ 1×1 卷积”。

小卷积替换大卷积时,常用小卷积核为 3×3 。LeCun等^[104]提出卷积神经检测网络LeNet-5架构,通过堆叠多个小卷积核组成的层,增加特征提取能力,证明了基于梯度学习的多层神经网络在模式识别中的优越性。

Simonyan等^[2]提出VGGNet网络架构,使用 3×3 小卷积核代替大卷积核,结合多个非线性激活层,使决策函数更具区分性。Li等^[105]提出shift-wise operator,将大卷积核分解为小卷积核并结合移位操作,通过剪枝实现稀疏组卷积,结合ghost和重参数化技术,模型精确率高达81.65%。但Shift-ConvNets在特定稀疏度下,时间消耗大。

除了常规 3×3 卷积核降低网络参数量之外,也可以通过 1×1 卷积核达到同样目的。He等^[106]提出RoShuNet轻量级卷积神经网络特征提取器,并采用聚合双组卷积模块(A-DGC)和 γ 加权洗牌模块(γ -WSM),引入瘦身策略。优化后模型的复杂度降低10.1%,模型尺寸减小26.4%,但应用瘦身策略后,模型准确率有所下降。

针对准确率降低,Tang等^[107]采用深度可分离卷积及批量归一化优化,得到轻量级一维卷积神经网络(Li-DtNet)模型。在嵌入式设备中,当信号采样频率不低于5 kHz时,故障识别和故障选线准确率均高于98%,适合轻量化部署。

小卷积核通过高效捕捉局部特征、减少过拟合并增强非线性表达,但需更多层扩大感受野,可能导致结构复杂和全局信息丢失,影响全局任务性能。

3.2.2 改进卷积结构

传统卷积因对每个通道独立应用核函数导致参数量大且冗余,深度可分离卷积、分组卷积和空洞卷积通过参数优化策略在减少计算量的同时保持特征提取能力,对该领域产生了重要影响并得到广泛应用。

Wang等^[108]提出一种轻量级的目标检测方法L-SSD,

该算法采用深度可分离卷积替换SSD网络中标准卷积,提出上采样特征融合模块(upsampling feature fusion module,UFFM)、局部-全局特征提取模块(local-global feature extraction module,LGFEM),改进加权双向特征金字塔(bidirectional feature pyramid,BiFPN),引入非对称空间注意力(asymmetric spatial attention,ASA)。深度可分离卷积将卷积操作分为深度卷积和逐点卷积两个部分,如图11所示,深度卷积对每个输入通道分别进行卷积,优化后模型参数和复杂度分别降低了85.9%和96.1%;逐点卷积则用于融合通道信息、扩展维度等。但BiFPN和ASA在提升精度的同时,检测速度有所下降。

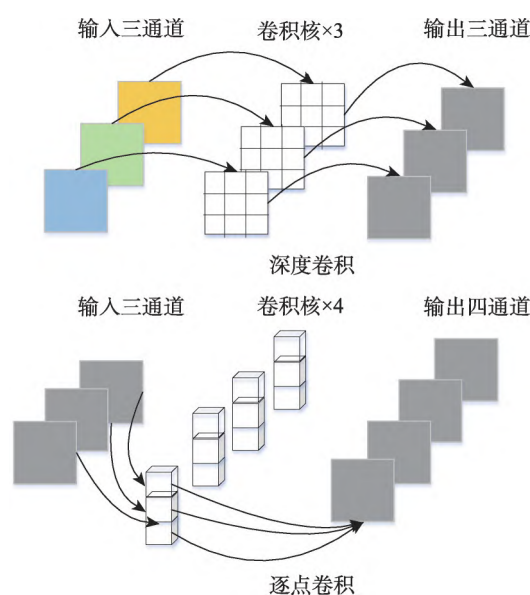


图11 深度可分离卷积过程

Fig.11 Depthwise separable convolution process

Krizhevsky等^[18]提出AlexNet网络,由于硬件限制引入分组卷积(group convolution)。利用多个GPU同时进行训练,最终合并结果以提高资源利用效率。分组卷积将输入沿通道分成若干组进行独立卷积处理,然后将其重组为输出特征,从而达到轻量化效果,如图12所示。简单分组卷积无法实现不同组之间信息交换。为

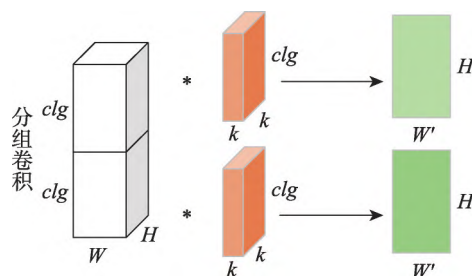


图12 分组卷积过程

Fig.12 Group convolution process

解决这个问题,Zhang等^[109]设计出一种移动级联组网络(SCGNet),采用移动分组卷积(shifting group convolution)的方法,它可以实现不同群体之间信息交换,堆叠一系列SCG块信息交换可达75.4%。但过多堆叠导致层数和训练难度增加。

除了深度可分离卷积和分组卷积,陈金荣等^[110]引入空洞卷积改进YOLOv5,提出YOLO-SSAR算法。结合前后置 1×1 卷积,优化感受野以提升多样化目标检测能力,设计多种不同比例的扩展卷积分支,提升多尺度特征提取能力,共享权重与残差连接,降低输入特征提取层的碎片化程度,相较于原模型检测精度提高了7.7%,这表明空洞卷积能够在减少参数量和计算量的基础上扩大图像的感受野。但推理时间较长,实时性略低。

3.2.3 总结分析

以上介绍的两种改进方式往往依赖于设计者的专业能力。研究者提出NAS,通过自动搜索出高效神经网络架构,并可能发现更为创新的网络结构,降低对设计者专业能力依赖。人工设计轻量化技术对比如表6所示。

3.3 深度学习模型压缩

目标检测精度为进一步提升,将会产生大量模型。因此,降低模型尺寸、计算量和内存占用的同时,保持模型性能,具有很高的研究价值。其中主要有剪枝、知识蒸馏、神经网络搜索(NAS)等模型压缩技术,直接减少模型参数量和结构大小。

3.3.1 剪枝

剪枝通过移除网络中一些权重(或神经元)来减少模型复杂度,如图13所示。剪枝后,模型大小和计算需求都会显著减少,同时仍然保持较高准确性。常见的一种剪枝流程包括对预训练剪枝,随后微调权重重新量化,以优化剪枝后的模型性能。

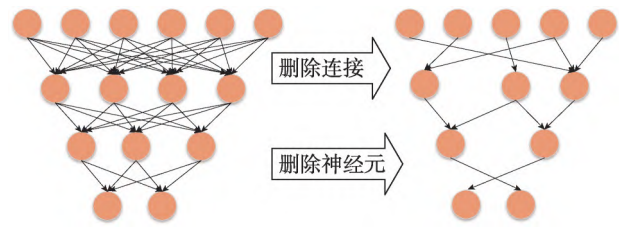


图13 模型剪枝过程

Fig.13 Model pruning process

OBD^[111]和OBS^[112]是经典非结构化剪枝方法,通过训练数据和网络结构信息来递推计算逆Hessian矩阵,基于Hessian矩阵评估损失函数的影响大小,更准确地删除不重要权重,从而减少网络复杂性。Shi等^[113]结合非结构化权重剪枝与非结构化神经元剪枝,重新参数化二进制掩码,设计能量惩罚项,得到新的剪枝框架。模型精度降低1.29%,计算量降低90.65%。但非结构化剪枝后模型不规则,处理器需要花时间处理被剪枝过的地方。

相比于修剪权重这种细粒度元素,结构化剪枝将滤波器、网络层这种粗粒度的元素作为裁剪对象,该方法重点研究不同结构的重要性。Luo等^[114]提出一种名为ThiNet的统一框架,通过使用下一层统计信息指导剪枝,从而将不重要滤波器移除,而不改变网络结构,显著减少了VGG-16和ResNet-50模型的参数量和计算量。但ThiNet的高压缩率下,模型的检测精度显著下降。

针对结构化剪枝可能改变网络层输出的特征图维度,Wang等^[115]提出基于强化学习的结构化剪枝(RL-Pruner)方法,利用Q-learning算法结合Monte Carlo采样策略,并在剪枝后进行知识蒸馏微调。该方法在CIFAR-100数据集上,VGG-19可压缩至60%通道稀疏度,ResNet-56准确率为93.71%。但RL-Pruner剪枝需多次微调,计算资源消耗大。Fang等^[116]提出一种通用且全自动的方法DepGraph(dependency graph),用于显式

表6 人工设计轻量化技术对比

Table 6 Comparison of manual design lightweight technology

类型	机制	优势	局限性	应用场景
调整卷积核大小	使用 3×3 卷积核替代大卷积核	参数减少,增加非线性层和网络深度,增强表达能力,提升特征学习能力	堆叠导致层数和训练难度增加,极端替代需多层堆叠	实时目标检测、移动端应用
	使用 1×1 卷积	显著减少参数和计算量,灵活调整通道维度,增强跨通道信息交互	无法捕捉空间相邻特征,过度降维丢失重要信息	特征融合、维度压缩/扩展、图像分类
改进卷积结构	深度可分离卷积	参数减少,计算效率高,适合移动端且保持特征表达能力	通道间信息组合受限,影响复杂任务性能	移动设备和嵌入式系统、实时检测
	分组卷积	参数减少,提高并行计算效率,促进模型轻量化	组间信息隔离,需额外操作,分组数需人工调优	大规模图像分类、目标检测、高分辨率图像分析
	空洞卷积	保持高分辨率下扩大感受野,多尺度特征融合,减少下采样信息丢失	扩张率过高导致局部信息丢失且增加计算复杂度	语义分割、医学图像分析、多尺度目标检测

建模层之间的依赖关系,并综合性地对分组耦合参数进行剪枝。但在部分配置下过度剪枝,导致模型性能和精度降低。

相较于权重,网络结构对模型预测更为重要。Liu等^[117]使用CIFAR-10、CIFAR-100和ImageNet数据集及VGG、ResNet和DenseNet网络架构评估多种结构化与非结构化剪枝方法。实验结果表明模型在结构化剪枝中表现优异,而非结构化剪枝未显著提升性能,修剪算法价值在于识别有效结构和执行网络结构搜索,而不是选择“重要”权重。

针对绝缘子缺陷检测较大的参数数量和计算量导致难以部署在边缘设备,亢洁等^[118]提出一种轻量化稀疏剪枝算法。通过偏移正则化稀疏训练策略,在剪枝过程中有效保留重要通道,避免过度剪枝。在CIFAR-10数据集上参数数量和计算量分别降低65.3%和66.0%。结合云边协同处理的特点,不仅提升模型的部署效率,还有助于改善复杂场景下的检测精度与泛化能力。但模型的推理时间较长,小样本学习能力弱。

3.3.2 知识蒸馏

知识蒸馏需要两种不同类型网络来分别作为学生模型和教师模型进行压缩,其过程如图14所示。教师模型作为训练好的模型,通过蒸馏训练将知识迁移至精简学生模型中,从而完成模型压缩。

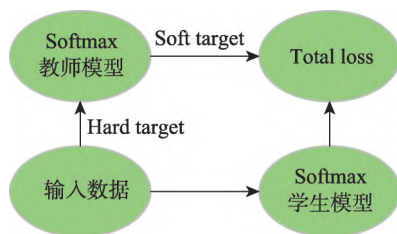


图14 知识蒸馏过程

Fig.14 Knowledge distillation process

2015年,Hiton^[119]将教师模型softmax层输出的logits作为知识传授给学生模型,这种基于响应(response-based)的知识定义思想,使得模型可以在较少数据和更高学习率下进行训练。Zhao等^[120]提出解耦知识蒸馏(double knowledge distillation, DKD)方法来进一步研究logits蒸馏,通过重新构建知识蒸馏(knowledge distillation, KD)损失为目标类知识蒸馏(target class knowledge distillation, TCKD)和非目标类知识蒸馏(non-target class knowledge distillation, NCKD)两个部分。TCKD通过二元逻辑蒸馏传递知识;NCKD只考虑非目标逻辑之间的知识,提升学生模型性能分别为74.79%、76.32%。相比经典KD, DKD在ImageNet数据集上准确率达71.7%,但基

于logit的蒸馏方法无法传递与目标定位相关的知识。

基于特征(feature-based)的知识指教师模型输出的特征图, Yang等^[121]提出一种新颖的跨图像关系知识蒸馏(cross-image relational knowledge distillation, CIRKD)方法。该方法引入像素对像素和像素对区域的蒸馏策略,通过内存机制充分探索结构化关系进行知识转移,来增强学生网络的语义分割性能。CIRKD与DeepLabV3和PSPNet相比,性能分别高出0.50%和0.73% mIoU。但引入内存机制和对比学习,增加了训练计算资源消耗,影响检测的实时性。

基于关系(relation-based)指学生模型学习不同层输出的特征图之间的关系, Park等^[122]引入一种通用的关系知识蒸馏(relational knowledge distillation, RKD)框架,通过两个简单有效的距离蒸馏和角度蒸馏损失来实现。在图像检索中,相比基线模型召回率可达60.67%;在图像分类中, RKD-DA与HKD结合准确率达74.66%;在少样本学习中, RKD学生模型准确率为98.64%。

在自动驾驶领域,目标检测的实时性和准确性尤为重要,王改华等^[123]提出基于知识蒸馏的轻量化Transformer目标检测模型(lightweight pooling Transformer, LPT)。通过设计通用知识蒸馏框架,提升模型的精度并降低计算复杂度。在MS COCO 2017数据集上计算复杂度降低29.6%,检测精度提升32.8%。使用双池化注意力机制,结合位置注意力和通道注意力的优势,提高特征提取效率和检测性能。但准确性相较于部分先进模型(如RT-DETR)略低。

3.3.3 神经网络搜索

神经网络搜索可以看作是训练超参数的超参数。随着可选超参数数量增加,搜索空间也会指数级增加,其过程如图15所示。Liu等^[124]提出一种新算法DARTS。传统方法通常依赖于进化或强化学习在离散、不可微分的搜索空间中进行搜索,而DARTS通过对架构表示进行连续化搜索,利用梯度下降实现高效搜索。2020年, Radosavovic等^[125]通过将初始AnyNet设计空间精简为RegNet设计空间。RegNet提供简单且快速的网络模型,以避免设计者经验不足导致搜索空间偏离问题。RegNet模型显著降低设计空间维度,大小缩减近10个数量级。但RegNet设计空间基于特定的网络结构,限

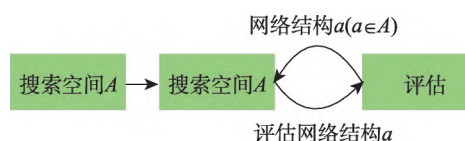


图15 神经网络搜索流程

Fig.15 Neural network search process

制在其他任务或输入分辨率中的泛化能力。

一般情况下,NAS都面临搜索成本太高的问题。这主要有两方面原因,一是搜索空间太大,二是需要训练的子网太多。为此Pham等^[126]提出一种名为高效神经架构搜索(efficient neural architecture search, ENAS)的方法。通过共享参数来提高效率,构建一个大型计算图,将所有可能的神经网络架构表示为该图的子图,并使用控制器RNN决定哪个子图被激活。实现了在不同搜索空间和领域中的灵活应用,并且ENAS错误率降低到2.89%。

近年来,针对搜索空间剪枝来减小搜索空间。Ci等^[127]提出一种名为神经搜索空间进化(neural search-space evolution, NSE)的方法。NSE从一个小的搜索空间子集开始,通过迭代优化和补充未遍历的操作,逐步扩展搜索空间,从而减少专门设计搜索空间的需求。但该方法训练时间长,计算资源要求高。Wang等^[128]提出一种LA-MCTS方法,该方法通过动态加深搜索树,选择和采样分割的搜索空间进行优化,减少计算量,加快推理速度。并利用SVM学习非线性决策边界,以提升性能。

搜索空间确定后,下一步就是从庞大的搜索空间中搜索优质网络结构。主要选用进化算法,强化学习等离散的搜索空间作为搜索策略。如Real等^[129]使用进化算法搜索网络结构。引入一种使用简化图作为DNA编码方法,将简化图转换为完整神经网络图进行训练和评估,实现对大规模模型架构搜索空间的探索。进化算法后的模型参数量减少为5.4 MB,准确率可达94.6%。Zoph等^[130]提出以强化学习自动化搜索最优架构。用RNN作控制器生成模型描述,将架构设计视为强化学习

问题,控制器在各时间步选择网络组件与参数,以验证其准确率为奖励信号训练RNN,引导其生成更优架构。

然而,这两种方法庞大的计算成本难以避免,为了降低搜索过程中计算成本使用基于梯度下降的搜索策略确定网络结构。如Xu等^[131]在DARTS基础上改进的可微架构搜索方法PC-DARTS。引入一种正则化机制,通过减少搜索和评估之间的剪枝差距,来减少网络空间冗余。PC-DARTS达到2.57%的错误率,显著优于DARTS,相比ProxylessNAS,PC-DARTS时间缩短为原来的1/2。

除了上述方法,Yang等^[132]将NAS引入雷达图像识别领域。采用知识蒸馏搜索策略,使用FSP蒸馏法,将NAS模型知识转移至计算复杂度更低的学生模型。优化后STIIARM模型参数量和计算复杂度分别减少96.05%和56.09%。Boutros等^[133]利用NAS结合多步知识蒸馏(multi-step KD),使用DARTS算法搜索特定于面部识别的架构。优化后的轻量级架构模型参数减少76.7%,同样实现模型的轻量化。

针对实时检测场景模型参数量多,Liu等^[134]提出一种基于知识蒸馏的NAS框架应用于驾驶分心识别的场景。引入了超级学生网络的概念,该网络通过加权和聚合所有候选架构,并利用教师网络进行指导,最终选择概率最大的候选项来构建学生网络。在USED数据集上参数量减少55%,精度提高2.91%。如表7总结了目标检测算法轻量化改进及结果对比。

由于传统遥感影像目标检测依赖专家经验且耗时久,杨军等^[135]提出基于NAS的自动化目标检测方法。使用逐路径采样结合进化搜索的策略,实现了高效的神

表7 目标检测算法轻量化改进

Table 7 Lightweight improvement of object detection algorithms

文献	相关改进	参数量/MB	计算量/GFLOPs	准确率/%	FPS
文献[80]	替换PReLU激活函数,引入SE注意力机制优化ShuffleNetv2	—	—	96.05	28.60
文献[81]	基于MBConv结构和SE模块改进ShuffleNetv2	2.17	0.28	98.50	—
文献[86]	使用GhostNet组件替代统卷积,引入GFB和GAL	0.63	0.59	92.90	—
文献[106]	轻量级RoShuNet特征提取器,A-DGC, γ -WSM,引入瘦身策略	5.78	0.32	94.03	140.00
文献[107]	使用一维DSC优化模型,引入批量归一化层	31.40	16.02	98.00	190.11
文献[108]	用DSC替代传统卷积,提出UFFM,LGFEM,改BiFPN,引入ASA	36.56	1.22	73.80	106.00
文献[113]	非结构化权重和神经元剪枝,重新参数化二进制掩码,设计能量惩罚项	—	8.50	81.90	—
文献[114]	设计ThiNet统一框架	5.05	1.72	91.40	37.30
文献[115]	Q-learning结合Monte Carlo采样策略,并在剪枝后进行知识蒸馏微调	0.86	—	93.71	71.25
文献[121]	提出CIRKD,同时结合传统任务损失和类概率KD损失	12.90	61.00	74.26	—
文献[125]	基于AnyNet设计空间精简为RegNet设计空间	107.80	31.70	80.50	318.00
文献[129]	使用进化算法自动发现神经网络架构	5.40	—	94.60	—
文献[132]	基于NAS和KD的轻量化CNN设计方法	2.09	0.81	96.30	3.30
文献[134]	使用渐进式学习构建教师网络,引导学生网络架构搜索并进行知识转移	3.28	2.25	95.64	2.23

经网络架构搜索,在 DIOR 和 RSOD 数据集上检测精度分别为 67.8% 和 85.5%。并显著减少了人工设计网络和调参的复杂性。相较于 NAS-FCOS 模型参数量和计算量分别降低 15.2% 和 6.8%,但在高分辨率遥感影像中对大型目标的检测精度有所下降。

3.3.4 量化

模型量化是一种将深度学习模型中浮点数参数转换为定点数参数的方法,旨在降低模型存储空间和计算资源消耗。目前量化方法根据量化是否参与模型训练主要分为训练后量化(post-training quantization, PTQ)和量化感知训练(quantization-aware training, QAT)。

传统 STE(straight through estimator)基础量化方法在 QAT 中优化网络导致不稳定性,影响准确性。Shin 等^[136]提出基于噪声代理的集成伪量化(noise proxy-based integrated pseudo-quantization, NIPQ)方法,该方法在不依赖 STE 的情况下实现稳定收敛。此方法首次将截断与离散化集成,不仅降低权重量化误差,还首次实现用于激活量化的 PQT。NIPQ 方法使计算量降低 47.78%,准确率提高 10.37%。但模型训练成本高,无法快速部署并应用到实际,因此与量化感知训练相比,训练后量化 PTQ 应用更为广泛。

为提高 PTQ 准确率,针对 PTQ 引入量化噪声问题, Liu 等^[137]提出 PD-Quant 的高精度后训练量化方法。引入预测差异损失(PD 损失)来优化量化参数,将 ResNet-18、RegNetX-600MF 准确率分别提高到 53.14%、40.67%。使用分布校正方法,以解决过拟合问题。在 PTQ 过程中,调整激活分布以符合批归一化层中保存的均值和方差,从而提高量化模型泛化能力。

这些 PTQ 方法针对已训练好的模型量化,微调时只需少量数据甚至无需数据,就能让模型部署到目标设备,实现简单且计算成本低。与之相反, QAT 在神经网络训练过程中模拟量化,虽要重新训练和调整超参数,但低 bit 时比 PTQ 更接近全精度效果。

为了实现高效、准确的纸板表面缺陷检测,陈俊杰等^[138]提出基于 FPGA 纸板缺陷检测系统。通过 QAT 对网络模型进行再训练和量化,在纸板缺陷数据集上参数量减少 74.9%,显著降低硬件资源消耗和计算复杂度。并设计了一种复用共享的硬件加速器架构,提高检测速度和能效。相较于原模型检测精度达到 88.67%,但量化后精度仍有略微的损失。

3.3.5 低秩分解

低秩分解是一种矩阵分解方法,主要包括奇异值分解 SVD(singular value decomposition)、张量分解等。旨在通过合并维数和施加低秩约束,来稀疏化矩阵使其变得更简洁,该过程如图 16 所示。

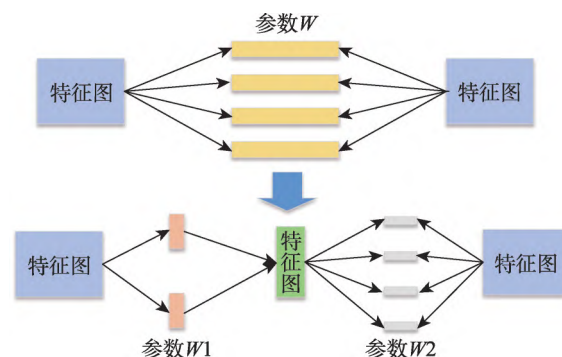


图 16 低秩分解过程

Fig.16 Low rank decomposition process

基于 SVD 模型压缩,一般是通过保留大矩阵的主要奇异值与对应奇异向量进行近似,以降低模型存储与计算复杂度。Yang 等^[139]采用 SVD 训练方法,使 DNN(deep neural network)达到低秩化,从而减少内存和计算负担,并提出正交性正则化和稀疏性诱导正则化技术,以在不显著降低模型性能的情况下实现模型压缩,为资源受限平台上 DNN 部署提供新方法。Mo 等^[140]提出一种加权双低秩分解(weighted double-low-rank decomposition, WDLRD)方法,通过为矩阵奇异值赋予不同的权重来保留织物图像中的重要特征。但 WDLRD 无法有效区分与背景对比度低的缺陷,深度网络效果低于浅层网络。

张量分解可以看作矩阵分解的泛化。Yin 等^[141]提出一种基于交替方向乘子法(alternating direction method of multipliers, ADMM)的系统性张量分解框架。该框架将 TT 分解模型压缩问题转化为一个带约束的非凸优化问题,并通过 ADMM 迭代求解。在此过程中,整个 DNN 模型逐渐展现低张量秩特性,然后再分解为 TT 格式并进行微调,最终获得高精度的 TT 格式 DNN 模型,但模型训练时间较长。Ahmed 等^[142]提出一种基于硬件感知的压缩方法。引入低秩分解结合压缩模式的概念,通过将网络层分组为不同压缩模式进行测试,此方法在保持准确性下降不超过 1%的前提下,实现模型压缩至原大小的 1/2。

为满足车载设备在资源有限情况下的高效计算,林德铝等^[143]提出一种基于低秩分解的参数压缩方法。结合 Tucker 分解压缩骨干和颈部的卷积层参数,实现保持精度的同时在 COCO 数据集上 YOLOv5-l 的参数量和计算量均减少 40%,在 OSDaR23 数据集上检测速度达到原模型的 1.5 倍。

3.3.6 总结分析

深度学习模型压缩的核心技术包括剪枝、知识蒸馏、神经网络搜索、量化和低秩分解。剪枝通过移除冗

余权重或结构降低模型复杂度,结构化剪枝更适配硬件但需平衡性能。知识蒸馏利用教师-学生模型传递知识高效保持精度,但存在内存和计算成本高。神经网络搜索自动化设计架构减少人工干预,但计算成本高。量化将浮点参数转为低比特整数,显著压缩存储,但复杂模型效果受限。低秩分解通过矩阵分解减少冗余、通用性强,但深度网络效果弱。

综上所述,将所提到模型压缩方法的特点优势以及局限性整理如表8所示。

3.4 总结分析

轻量级网络通过优化骨干、颈部和头部网络结构实现高效推理。轻量级卷积设计方面,调整卷积核大小和改进卷积结构,有效减少冗余计算,同时保持特征提取能力。模型压缩技术通过剪枝移除冗余权重或结构,知识蒸馏实现教师-学生模型的知识迁移,NAS自动化设计高效架构,量化将浮点参数转为低比特整数,低秩分解通过矩阵分解减少存储与计算量。这些方法在降低模型复杂度的同时,权衡精度损失与硬件适配性,共同推动目标检测算法在资源受限场景下的实用化部署。

4 轻量化目标检测应用

4.1 小目标检测

小目标检测因其低分辨率、弱语义特征及复杂背景干扰,始终是检测技术的难点。常规目标检测却因参数量大和计算复杂度高,难以适配无人机、嵌入式传感器等资源受限场景。研究者提出针对小目标检测的轻量化算法。

SSD通常使用轻量化的骨干网络作为特征提取网

络,提高了对小目标的检测能力。如Yang等^[144]采用FasterNet替换SSD原有的VGG16主干网络,并引入特征增强模块FEM和深度特征融合模块DFFM,提出FasterNet-SSD模型。改进后的模型在减少计算复杂度的同时提高了检测精度。

CenterNet小目标检测存在检测框质量差、模型参数和计算量大等缺点。岳永恒等^[145]提出轻量化的CenterNet。通过引入轻量化MobileNetv3网络,优化注意力机制、激活函数和特征融合模块实现轻量化。相较于原模型参数量减少72.7%,检测精度提升12%,但检测速度有所降低。

针对小目标检测存在的漏检、误检、精度与速度不平衡等问题,廖宁生等^[146]提出轻量化的CM-YOLOv8s。通过引入通道特征优化SPPF模块和重构颈部结构,提升模型定位和提取能力。在VisDrone2019数据集上参数量降低70.2%,mAP提升4.9%。设计动态检测头(DRHead)实现感受野动态调整,平衡检测精度与速度。但由于浅层特征图尺寸较大,模型计算量增加。

为了降低模型计算量,林永洪等^[147]提出轻量化的UD-DETR模型。在RT-DETR模型针对多尺度小目标异物检测的基础上,结合SFDCConv改进主干网络卷积计算,增强多尺度特征提取。在科大讯飞检测挑战赛数据集上参数量和计算量分别降低44.7%和30.8%。优化路径聚合网络UDFPN增强小目标特征提取能力,显著提升小目标检测精度和性能。但在实时场景中检测的实时性有待提升。

为了解决上述问题,刘延芳等^[148]基于多尺度多深度特征提取(multi-scalar & multi-depth feature extraction,

表8 模型压缩方法对比

Table 8 Comparison of model compression methods

压缩方法	机制	文献方法	优势	局限性	应用场景
剪枝	移除模型中不重要的权重或神经元,减少模型的参数量和计算复杂度	RL-Pruner ^[115] DepGraph ^[116]	非均匀分配稀疏度,高效处理复杂网络,多任务场景通用	计算成本高,部分模型因结构限制性能下降	适用于边缘设备部署
知识蒸馏	训练一个学生模型学习教师模型的输出,从而实现模型的轻量化,保持较高的性能	DKD ^[120] CIRKD ^[121]	高效训练,捕捉全局语义关系,提升分割精度,泛化能力强	内存和计算成本高,定位知识传递不足,超参数调整需研究	图像分类、语义分割
神经网络搜索	算法自动设计并优化神经网络结构,以满足特定性能需求和计算资源限制	NSE ^[127] PC-DARTS ^[131]	内存效率高,搜索速度快,稳定性强,减少对超参数的敏感性	计算复杂度较高,需多次迭代,部分通道采样信息丢失,影响准确性	资源受限环境、大规模数据集
量化	减少模型参数的表示精度,将浮点数转换为低比特数的整数表示	NIPQ ^[136] PD-Quant ^[137]	低bit仍保持高精度,无需训练数据,缓解过拟合,提升泛化能力	计算复杂度较高,依赖梯度下降,对复杂模型效果受限	适用于边缘设备、移动设备上的实时推理
低秩分解	将模型中的权重矩阵分解为多个低秩矩阵的乘积,从而减少参数量和计算复杂度	改进奇异值分解 ^[139] 张量分解 ^[141]	相同精度下计算量低,通用性强,压缩后精度损失小	深度网络效果低于浅层网络,模型复杂训练时间较长	图像分类、语义分割,边缘设备部署

MMFE)网络提出(real-time small target detection, RTSTD)算法。该算法通过裁剪图像和重叠向量技术,设计轻量化的MMFE网络,并重新定义正负样本来增强背景与目标的区分能力。在无人机遥感图像中,对于小目标检测的实时性表现出色。

4.2 工业检测

基于机器视觉的工业检测包括缺陷检测、目标定位、字符识别等一系列替代人眼,实现高精度和高速度的检测。为了应对在复杂工业场景中硬件性能限制、效率成本等问题,研究者提出多种针对工业检测的轻量化算法。

针对钢材表面缺陷检测效率低和不易部署在移动端设备,周赵轩等^[149]提出轻量化ZZX-YOLO算法。通过设计卷积模块ZZXConv和轻量化特征提取模块ZZX_CSPC,增强模型非线性表达能力和减少冗余信息。在工业级GC10-DET钢材表面缺陷数据集上参数量和计算量分别下降8.52%和9.27%。引入K-means++重新设计先验框,增强稳定性和泛化能力,相较于原模型检测精度提升7.7%。胡玮等^[150]采用轻量级网络Faster-Net替代DarkNet,并对YOLOv7-Tiny进行通道剪枝,同样实现了模型轻量化。

对于钢材表面缺陷检测精度低、模型复杂,梁礼明等^[151]提出EHH-YOLO。将颈部网络H-C2f模块替换C2f结构,减少特征冗余和计算负载。在NEU-DET钢材缺陷检测数据集上参数量和计算量分别减低36.9%和45.8%。采用HWD模块改善缺陷中漏检、误检问题,相较于原模型检测精度提升6.9%,但下采样操作扩大特征图,导致模型计算量增加。

为了减少模型计算量和复杂度,王春梅等^[152]提出轻量的YOLOv8-VSC。使用轻量级VanillaNet作为骨干网络,减少分支结构降低模型复杂度。在NEU-DET数据集上,参数量和计算量分别降低42.5%和38.3%。引入轻量级上采样算子CARAFE优化特征融合,相较于原模型检测精度提升1.2%,但CARAFE中内容感知重组过程会保留较多的空间信息。

为了解决上述问题,刚帅等^[153]使用轻量化Shuffle-Netv2作为主干网络,既降低模型复杂度,又避免了组卷积引入的信息传递瓶颈。在NEU-DET钢材表面缺陷数据集上模型参数量和体积分别降低39.2%和42.9%。加入BiFPN促进特征提取和信息融合,采用双卷积核模块(DualConv)替代C2f中的卷积层,相较于原模型计算量降低41.5%,mAP提升0.3%。

4.3 自动驾驶

自动驾驶的环境感知高度依赖目标检测技术,其需

实时识别道路中的车辆、行人、交通标志及复杂障碍物。然而,当前算法面临实时性、小目标漏检等多重局限。因此,亟需通过轻量化目标检测算法,为其提供可靠性与安全性。

针对车辆检测参数量和计算量大、识别精度低、小目标检测难,许晓阳等^[154]提出KD-YOLO-DW算法。采用轻量化GhostNetv2网络,减少计算复杂度的同时提高特征提取。结合DSC设计轻量化ELAN-DW模块,在FLIR数据集上参数量降低36.1%。并改进动态非单调聚焦机制的WIoU损失函数,以平衡小目标和模糊目标检测精度,改进后检测精度提升3.27%。

为应对交通标志检测小目标精度低和模型复杂度高,罗向龙等^[155]采用GhostBottleneckv2结合GSConv模块,降低模型复杂度提升收敛速度。在TT100K数据集上参数量和计算量分别降低16.8%和2.5%。使用WIoUv3回归损失函数动态优化小目标的损失权重,相较于原模型检测精度提升2.4%,但在复杂环境下的检测性能存在不足。

对于复杂背景下检测性能不足,刘菲等^[156]提出BMGE-YOLOv5s算法。采用BoTNet(bottleneck Transformer network)骨干网络降低背景差异性小对检测精度的影响,结合MPDIoU损失函数提升边界框定位,在CCTSD2017数据集上检测精度和速度分别提升2.5%和8.3%。并引入C3GBneckv2优化颈部网络,降低模型参数量,显著提升检测性能。

为解决行人检测难以平衡速度和精度问题,王泽宇等^[157]提出轻量级MER-YOLO算法。采用MobileViT作为主干网络,增强特征提取能力,减少冗余特征加快训练速度。在Crowd Human数据集上参数量和计算量分别降低56%和35%。结合EMA注意力机制并加小目标检测头,相较于原模型检测精度提升4.5%,在复杂场景中表现出更强的鲁棒性和泛化性能。

4.4 总结分析

轻量化目标检测在小目标检测、工业检测和自动驾驶等领域取得显著进展,通过优化网络结构、引入高效特征提取模块,有效平衡了模型性能与计算资源消耗。在小目标检测中,研究聚焦于增强多尺度特征融合、轻量化骨干网络及改进损失函数,在无人机遥感、复杂场景异物检测等场景中实现了检测精度与速度的提升。工业检测领域则通过设计专用卷积模块、优化特征金字塔及数据增强,在降低参数的同时提高缺陷检测效率,但部分模型存在小目标漏检或泛化能力不足的问题。自动驾驶场景中,轻量化算法结合Transformer、知识蒸馏及动态损失函数,在实时性与复杂环境鲁棒性上取得

突破,但极端天气下的检测稳定性和边缘设备部署效率仍需进一步优化。

总体而言,轻量化技术通过模块化设计、动态机制及硬件友好型算子,为资源受限场景提供了高效解决方案。

5 总结与展望

本文围绕基于深度学习的目标检测算法轻量化展开总结。介绍了目标检测算法及轻量化评价指标。重点阐述轻量化关键技术,轻量级网络创新结构;轻量级卷积设计优化从调整卷积核大小和改进结构着手;模型压缩含剪枝、知识蒸馏等多种方法。未来可从多学科融合、硬件适配、边缘设备部署等方面展望。

5.1 多学科技术融合

将深度学习与其他学科技术如计算机视觉、图像处理、模式识别等进一步融合,充分发挥各技术优势,提升目标检测算法性能。Che等^[158]将计算机视觉与机器人控制结合,展示机器人在复杂环境中自适应能力,使其能够通过实时视觉数据分析进行自主导航和目标识别。该机器人利用图像识别和特征提取技术实现高精度分类,提高检测处理效率并减少人工成本。此外,还提出一种实时监测与自动识别系统,通过高清摄像头和深度学习算法对区域进行监控,能够快速识别图像并发出预警,从而提升检测效率和准确性,减少环境影响。

除此之外,与量子计算、生物识别等技术融合,如熊其冰等^[159]将量子计算融入经典卷积神经网络,以提高模型的计算效率,王东^[160]将深度学习模型从生物特征面部图像生成非个人数字特征,提高以人脸图像作为生物特征识别和信息认证系统的安全性,开拓深度学习在工业自动化、医疗保健和环境保护等领域的应用。

5.2 硬件架构设计优化

针对轻量化目标检测算法,设计专门硬件架构,以更好支持和加速算法运行。如开发针对深度可分离卷积、分组卷积等特殊卷积结构的硬件加速器^[161],提高计算效率,减少硬件资源消耗,使算法在资源受限设备上更高效运行。Samanta等^[162]提出对边缘设备深度学习硬件加速器的设计,通过基于边缘计算的硬件加速器设计、数据流重新配置与功耗管理以及近似计算和低阈值计算等方法。TinyVer芯片通过数据流重新配置和片上功率管理实现多模式应用和智能感知,而自补偿加速器(SeCAs)通过使用通用精度可配置加法器和低功耗乘法器来减少数据流误差,提高计算精度。旨在提高边缘设备上深度学习算法的效率、功耗和准确性。

5.3 边缘设备部署策略

随着物联网和边缘计算发展,将轻量化目标检测算

法部署在边缘设备上具有重要意义。研究如何优化算法使其适应边缘设备的计算能力、存储资源和功耗限制,同时保证检测性能,实现实时、高效的目标检测应用。如在智能家居、智能安防等领域的应用。Sun等^[163]针对大语言模型在边缘设备上部署,开发了新型AI边缘设备,支持低位量化计算加速,并采用PTQ和QAT技术,有效降低模型计算复杂性和内存占用。此外,还探索有效的Transformer架构设计,进一步减少大语言模型参数量和推理时间,使其能够在边缘设备上高效运行。

参考文献:

- [1] HE C X, TAN S B, ZHAO J, et al. Efficient and lightweight neural network for hard hat detection[J]. Electronics, 2024, 13(13): 2507.
- [2] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[EB/OL]. [2025-02-10]. <https://arxiv.org/abs/1409.1556>.
- [3] CHEN F H, LI S L, HAN J L, et al. Review of lightweight deep convolutional neural networks[J]. Archives of Computational Methods in Engineering, 2024, 31(4): 1915-1937.
- [4] MITTAL P. A comprehensive survey of deep learning-based lightweight object detection models for edge devices[J]. Artificial Intelligence Review, 2024, 57(9): 242.
- [5] 董甲东, 郭庆虎, 陈琳, 等. 深度学习中单阶段金属表面缺陷检测算法优化综述[J]. 计算机工程与应用, 2025, 61(4): 72-89.
- [6] DONG J D, GUO Q H, CHEN L, et al. Review on optimization algorithms for one-stage metal surface defect detection in deep learning[J]. Computer Engineering and Applications, 2025, 61(4): 72-89.
- [7] KAUR R, SINGH S. A comprehensive review of object detection with deep learning[J]. Digital Signal Processing, 2023, 132: 103812.
- [8] CAO Y, MIAO Q G, LIU J C, et al. Advance and prospects of AdaBoost algorithm[J]. Acta Automatica Sinica, 2013, 39(6): 745-758.
- [9] HEARST M A, DUMAIS S T, OSUNA E, et al. Support vector machines[J]. IEEE Intelligent Systems and Their Applications, 1998, 13(4): 18-28.
- [10] AHMED A, JALAL A, KIM K. Region and decision tree-based segmentations for multi-objects detection and classification in outdoor scenes[C]//Proceedings of the 2019 International Conference on Frontiers of Information Technology. Piscataway: IEEE, 2019: 209-214.
- [11] VIOLA P, JONES M. Rapid object detection using a boosted

- cascade of simple features[C]//Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2001.
- [11] ALI S, HANZLA M, RAFIQUE A A. Vehicle detection and tracking from UAV imagery via cascade classifier[C]//Proceedings of the 2022 24th International Multitopic Conference. Piscataway: IEEE, 2022: 1-6.
- [12] LINDBERGER T. Scale invariant feature transform[J]. Scholarpedia, 2012, 7(5): 10491.
- [13] 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. Piscataway: IEEE, 2005: 886-893.
- [14] TANG Z T, ZHANG Z M, CHEN W, et al. An SIFT-based fast image alignment algorithm for high-resolution image [J]. IEEE Access, 2023, 11: 42012-42041.
- [15] BARROS B, CONDE B, CABALEIRO M, et al. Design and testing of a decision tree algorithm for early failure detection in steel truss bridges[J]. Engineering Structures, 2023, 289: 116243.
- [16] ZHANG L, XU W Y, SHEN C, et al. Vision-based on-road nighttime vehicle detection and tracking using improved HOG features[J]. Sensors, 2024, 24(5): 1590.
- [17] ABBAS Q, ALBALAWI T S, PERUMAL G, et al. Automatic face recognition system using deep convolutional mixer architecture and AdaBoost classifier[J]. Applied Sciences, 2023, 13(17): 9880.
- [18] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks[C]//Advances in Neural Information Processing Systems 25, 2012.
- [19] GIRSHICK R, DONAHUE J, DARRELL T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2014: 580-587.
- [20] 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.
- [21] GIRSHICK R. Fast R-CNN[C]//Proceedings of the 2015 IEEE International Conference on Computer Vision. Piscataway: IEEE, 2015: 1440-1448.
- [22] REN S Q, HE K M, GIRSHICK R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[C]//Advances in Neural Information Processing Systems 28, 2015.
- [23] HE K M, GKIOXARI G, DOLLÁR P, et al. Mask R-CNN [C]//Proceedings of the 2017 IEEE International Conference on Computer Vision. Piscataway: IEEE, 2017: 2980-2988.
- [24] CAI Z W, VASCONCELOS N. Cascade R-CNN: delving into high quality object detection[C]//Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 6154-6162.
- [25] SUN P Z, ZHANG R F, JIANG Y, et al. Sparse R-CNN: end-to-end object detection with learnable proposals[C]//Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2021: 14449-14458.
- [26] XU X K, FENG Z J, CAO C Q, et al. An improved swin transformer-based model for remote sensing object detection and instance segmentation[J]. Remote Sensing, 2021, 13(23): 4779.
- [27] LIN T Y, DOLLÁR P, GIRSHICK R, et al. Feature pyramid networks for object detection[C]//Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2017: 2117-2125.
- [28] GHIASI G, LIN T Y, LE Q V. NAS-FPN: learning scalable feature pyramid architecture for object detection[C]//Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2019: 7036-7045.
- [29] 杨扬, 唐晓芬. ResFPN: 扩增实际感受野和改进 FPN 的多尺度目标检测方法[J]. 计算机工程与应用, 2025, 61(10): 247-257.
- YANG Y, TANG X F. ResFPN: multi-scale object detection algorithm for expanding actual receptive field and improving FPN[J]. Computer Engineering and Applications, 2025, 61(10): 247-257.
- [30] YU Y, YANG X, LI Q, et al. H2RBox-v2: incorporating symmetry for boosting horizontal box supervised oriented object detection[C]//Advances in Neural Information Processing Systems 36, 2023: 59137-59150.
- [31] 冷岳峰, 刘正, 徐宝祎, 等. 改进 Faster R-CNN 的钢材表面缺陷检测[J]. 机械科学与技术, 2025, 44(1): 75-83.
- LENG Y F, LIU Z, XU B Y, et al. Detection of steel surface defect based on improved faster R-CNN[J]. Mechanical Science and Technology for Aerospace Engineering, 2025, 44(1): 75-83.
- [32] REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: unified, real-time object detection[C]//Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern

- Recognition. Piscataway: IEEE, 2016: 779-788.
- [33] REDMON J, FARHADI A. YOLO9000: better, faster, stronger[C]//Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2017: 7263-7271.
- [34] REDMON J, FARHADI A. YOLOv3: an incremental improvement[EB/OL]. [2025-02-10]. <https://arxiv.org/abs/1804.02767>.
- [35] BOCHKOVSKIY A, WANG C Y, LIAO H Y M. YOLOv4: optimal speed and accuracy of object detection[EB/OL]. [2025-02-10]. <https://arxiv.org/abs/2004.10934>.
- [36] ZAIDI S S A, ANSARI M S, ASLAM A, et al. A survey of modern deep learning based object detection models[J]. Digital Signal Processing, 2022, 126: 103514.
- [37] GE Z, LIU S, WANG F, et al. YOLOX: exceeding YOLO series in 2021[EB/OL]. [2025-02-10]. <https://arxiv.org/abs/2107.08430>.
- [38] LI C, LI L, JIANG H, et al. YOLOv6: a single-stage object detection framework for industrial applications[EB/OL]. [2025-02-10]. <https://arxiv.org/abs/2209.02976>.
- [39] WANG C Y, BOCHKOVSKIY A, LIAO H M. YOLOv7: trainable bag-of-freebies sets new state-of-the-art for real-time object detectors[C]//Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2023: 7464-7475.
- [40] YASEEN M. What is YOLOv9: an in-depth exploration of the internal features of the next-generation object detector [EB/OL]. [2025-02-10]. <https://arxiv.org/abs/2409.07813>.
- [41] WANG C Y, YE H I, MARK LIAO H Y. YOLOv9: learning what you want to learn using programmable gradient information[C]//Proceedings of the 18th European Conference on Computer Vision. Cham: Springer, 2024: 1-21.
- [42] WANG A, CHEN H, LIU L, et al. YOLOv10: real-time end-to-end object detection[C]//Advances in Neural Information Processing Systems 37, 2025: 107984-108011.
- [43] KHANAM R, HUSSAIN M. YOLOv11: an overview of the key architectural enhancements[EB/OL]. [2025-02-10]. <https://arxiv.org/abs/2410.17725>.
- [44] LIU W, ANGUELOV D, ERHAN D, et al. SSD: single shot multibox detector[C]//Proceedings of the 14th European Conference on Computer Vision. Cham: Springer, 2016: 21-37.
- [45] FU C Y, LIU W, RANGA A, et al. DSSD: deconvolutional single shot detector[EB/OL]. [2025-02-10]. <https://arxiv.org/abs/1701.06659>.
- [46] JEONG J, PARK H, KWAK N. Enhancement of SSD by concatenating feature maps for object detection[EB/OL]. [2025-02-10]. <https://arxiv.org/abs/1705.09587>.
- [47] LI Z X, ZHOU F Q. FSSD: feature fusion single shot multibox detector[EB/OL]. [2025-02-10]. <https://arxiv.org/abs/1712.00960>.
- [48] 董一兵, 曾辉, 侯少杰. LMUAV-YOLOv8: 低空无人机视觉目标检测轻量化网络[J]. 计算机工程与应用, 2025, 61(3): 94-110.
- DONG Y B, ZENG H, HOU S J. LMUAV-YOLOv8: light-weight network for object detection in low-altitude UAV vision[J]. Computer Engineering and Applications, 2025, 61(3): 94-110.
- [49] 雷景生, 章志豪, 钱小鸿, 等. 改进YOLOX的轻量级多方向车牌检测算法[J]. 计算机工程与应用, 2025, 61(4): 230-240.
- LEI J S, ZHANG Z H, QIAN X H, et al. Improved light-weight multi-directional license plate detection algorithm of YOLOX[J]. Computer Engineering and Applications, 2025, 61(4): 230-240.
- [50] LIN T Y, GOYAL P, GIRSHICK R, et al. Focal loss for dense object detection[C]//Proceedings of the 2017 IEEE International Conference on Computer Vision. Piscataway: IEEE, 2017: 2980-2988.
- [51] DUAN K W, BAI S, XIE L X, et al. CenterNet: keypoint triplets for object detection[C]//Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2019: 6568-6577.
- [52] TIAN Z, SHEN C H, CHEN H, et al. FCOS: a simple and strong anchor-free object detector[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 44(4): 1922-1933.
- [53] WU B C, WAN A, IANDOLA F, et al. SqueezeDet: unified, small, low power fully convolutional neural networks for real-time object detection for autonomous driving[C]//Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops. Piscataway: IEEE, 2017: 446-454.
- [54] TAN M X, PANG R M, LE Q V. EfficientDet: scalable and efficient object detection[C]//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2020: 10778-10787.
- [55] CARION N, MASSA F, SYNNAEVE G, et al. End-to-end object detection with transformers[C]//Proceedings of the 16th European Conference on Computer Vision. Cham: Springer, 2020: 213-229.
- [56] ZHU X, SU W, LU L, et al. Deformable DETR: deformable transformers for end-to-end object detection[EB/OL]. [2025-02-10]. <https://arxiv.org/abs/2010.04159>.

- [57] DAI X Y, CHEN Y P, YANG J W, et al. Dynamic DETR: end-to-end object detection with dynamic attention[C]//Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2021: 2968-2977.
- [58] MENG D P, CHEN X K, FAN Z J, et al. Conditional DETR for fast training convergence[C]//Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2021: 3631-3640.
- [59] LIU S, LI F, ZHANG H, et al. DAB-DETR: dynamic anchor boxes are better queries for DETR[EB/OL]. [2025-02-10]. <https://arxiv.org/abs/2201.12329>.
- [60] LI F, ZHANG H, LIU S L, et al. DN-DETR: accelerate DETR training by introducing query DeNoising[C]//Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2022: 13609-13617.
- [61] ZHANG M Y, SONG G L, LIU Y, et al. Decoupled DETR: spatially disentangling localization and classification for improved end-to-end object detection[C]//Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2023: 6578-6587.
- [62] ZHANG J Y, HUANG J X, LUO Z P, et al. DA-DETR: domain adaptive detection transformer with information fusion[C]//Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2023: 23787-23798.
- [63] LI F, ZENG A L, LIU S L, et al. Lite DETR: an interleaved multi-scale encoder for efficient DETR[C]//Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2023: 18558-18567.
- [64] ZHENG D H, DONG W H, HU H L, et al. Less is more: focus attention for efficient DETR[C]//Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2023: 6651-6660.
- [65] ZHAO Y A, LV W Y, XU S L, et al. DETRs beat YOLOs on real-time object detection[C]//Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2024: 16965-16974.
- [66] ZHAO C Y, SUN Y F, WANG W H, et al. MS-DETR: efficient DETR training with mixed supervision[C]//Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2024: 17027-17036.
- [67] 王鹏九, 龚俊斌, 罗威, 等. 基于改进 Deformable DETR 的水面目标检测[J]. 中国舰船研究, 2025, 20(3): 305-317.
- WANG P J, GONG J B, LUO W, et al. Detection of water surface targets based on improved Deformable DETR[J]. Chinese Journal of Ship Research, 2025, 20(3): 305-317.
- [68] 冯永安, 张紫扬, 张旭. 双重细化门控自适应融合的道路裂缝检测算法[J/OL]. 计算机科学与探索[2025-04-05]. <https://kns.cnki.net/kcms/detail/11.5602.TP.20250320.1614.006.html>.
- FENG Y A, ZHANG Z Y, ZHANG X. Dual refinement gate control adaptive fusion algorithm for road crack detection [J/OL]. Journal of Frontiers of Computer Science and Technology[2025-04-05]. <https://kns.cnki.net/kcms/detail/11.5602.TP.20250320.1614.006.html>.
- [69] LIU F L, ZHENG Q H, TIAN X Y, et al. Rethinking the multi-scale feature hierarchy in object detection transformer (DETR)[J]. Applied Soft Computing, 2025, 175: 113081.
- [70] CHEN H, TANG C F, HU X L. DHS-DETR: efficient DETRs with dynamic head switching[J]. Computer Vision and Image Understanding, 2024, 248: 104106.
- [71] HOWARD A G, ZHU M, CHEN B, et al. MobileNets: efficient convolutional neural networks for mobile vision applications[EB/OL]. [2025-02-15]. <https://arxiv.org/abs/1704.04861>.
- [72] SANDLER M, HOWARD A, ZHU M L, et al. MobileNetV2: inverted residuals and linear bottlenecks[C]//Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 4510-4520.
- [73] HOWARD A, SANDLER M, CHEN B, et al. Searching for MobileNetV3[C]//Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2019: 1314-1324.
- [74] TAN M X, CHEN B, PANG R M, et al. MnasNet: platform-aware neural architecture search for mobile[C]//Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2019: 2815-2823.
- [75] QIN D F, LEICHNER C, DELAKIS M, et al. MobileNetV4: universal models for the mobile ecosystem[C]//Proceedings of the 18th European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2025: 78-96.
- [76] 肖振久, 杨晓迪, 魏宪, 等. 改进的轻量型网络在图像识别上的应用[J]. 计算机科学与探索, 2021, 15(4): 743-753.
- XIAO Z J, YANG X D, WEI X, et al. Improved lightweight network in image recognition[J]. Journal of Frontiers of Computer Science and Technology, 2021, 15(4): 743-753.
- [77] JUNAYED M S, ISLAM M B, IMANI H, et al. PDS-Net: a novel point and depth-wise separable convolution for real-time object detection[J]. International Journal of Multimedia Information Retrieval, 2022, 11(2): 171-188.
- [78] ZHANG X Y, ZHOU X Y, LIN M X, et al. ShuffleNet:

- an extremely efficient convolutional neural network for mobile devices[C]//Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 6848-6856.
- [79] MA N N, ZHANG X Y, ZHENG H T, et al. ShuffleNet V2: practical guidelines for efficient CNN architecture design[C]//Proceedings of the 15th European Conference on Computer Vision. Cham: Springer, 2018: 122-138.
- [80] YANG H J, LIU J X, MEI G M, et al. Research on real-time detection method of rail corrugation based on improved ShuffleNet V2[J]. Engineering Applications of Artificial Intelligence, 2023, 126: 106825.
- [81] ZHANG Y X, XIE W, YU X Y. Design and implementation of liveness detection system based on improved ShuffleNet V2[J]. Signal, Image and Video Processing, 2023, 17(6): 3035-3043.
- [82] IANDOLA F N, HAN S, MOSKEWICZ M W, et al. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5 MB model size[EB/OL]. [2025-02-15]. <https://arxiv.org/abs/1602.07360>.
- [83] HAN K, WANG Y H, TIAN Q, et al. GhostNet: more features from cheap operations[C]//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2020: 1577-1586.
- [84] TANG Y, HAN K, GUO J, et al. GhostNetV2: enhance cheap operation with long-range attention[C]//Advances in Neural Information Processing Systems 35, 2022: 9969-9982.
- [85] LIU Z, HAO Z, HAN K, et al. GhostNetV3: exploring the training strategies for compact models[EB/OL]. [2025-02-15]. <https://arxiv.org/abs/2404.11202>.
- [86] ARDIYANTO I. Edge devices-oriented surface defect segmentation by GhostNet fusion block and global auxiliary layer [J]. Journal of Real-Time Image Processing, 2023, 21(1): 13.
- [87] 王迎龙, 孙备, 丁冰, 等. BG-YOLO: 复杂大视场下低慢小无人机目标检测方法[J]. 仪器仪表学报, 2025, 46(2): 255-266.
- WANG Y L, SUN B, DING B, et al. BG-YOLO: a low-altitude slow-moving small UAV targets detection method in complex large field of view[J]. Chinese Journal of Scientific Instrument, 2025, 46(2): 255-266.
- [88] TAN M, LE Q. EfficientNet: rethinking model scaling for convolutional neural networks[C]//Proceedings of the 36th International Conference on Machine Learning, 2019: 6105-6114.
- [89] TAN M, LE Q. EfficientNetV2: smaller models and faster training[C]//Proceedings of the 38th International Conference on Machine Learning, 2021: 10096-10106.
- [90] 卜晓燕, 张宪法, 李明慧, 等. 基于改进 EfficientDet 的飞机蒙皮缺陷检测方法[J]. 航空制造技术, 2025, 68(5): 68-75.
- BU X Y, ZHANG X F, LI M H, et al. Aircraft skin defect detection method based on improved EfficientDet[J]. Aeronautical Manufacturing Technology, 2025, 68(5): 68-75.
- [91] 田栋, 魏霞, 袁杰. 基于改进 YOLOv5 轻量化的车辆目标检测算法[J]. 计算机应用与软件, 2024, 41(12): 240-246.
- TIAN D, WEI X, YUAN J. Vehicle target detection algorithm based on improved YOLOv5 lightweight[J]. Computer Applications and Software, 2024, 41(12): 240-246.
- [92] CHEN J R, KAO S H, HE H, et al. Run, don't walk: chasing higher FLOPS for faster neural networks[C]//Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2023: 12021-12031.
- [93] GUO A, SUN K Q, ZHANG Z Y. A lightweight YOLOv8 integrating FasterNet for real-time underwater object detection [J]. Journal of Real-Time Image Processing, 2024, 21(2): 49.
- [94] DONG X, LI D, FANG J D. FCCD-SAR: a lightweight SAR ATR algorithm based on FasterNet[J]. Sensors, 2023, 23(15): 6956.
- [95] MA X, DAI X Y, BAI Y, et al. Rewrite the stars[C]//Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2024: 5694-5703.
- [96] CHU Y Q, YU X Y, RONG X W. A lightweight strip steel surface defect detection network based on improved YOLOv8[J]. Sensors, 2024, 24(19): 6495.
- [97] 古莹奎, 叶彪彪, 郭明健, 等. 基于改进 RT-DETR 的饼干包装外观缺陷快速检测[J]. 食品与机械, 2025, 41(2): 234-241.
- GU Y K, YE B B, GUO M J, et al. Rapid detection method of biscuit packaging appearance defects based on improved RT-DETR[J]. Food & Machinery, 2025, 41(2): 234-241.
- [98] 葛雯, 邵钰琦, 屈乐乐. 面向遥感图像的轻量化小目标检测算法研究[J]. 电子测量技术, 2025, 48(4): 118-127.
- GE W, SHAO Y Q, QU L L. Research on lightweight small target detection algorithm for remote sensing images[J]. Electronic Measurement Technology, 2025, 48(4): 118-127.
- [99] 沈筹, 张磊, 张宇翔, 等. 基于改进 YOLOv8n 的轻量化分心驾驶行为检测方法[J]. 电子测量技术, 2024, 47(24): 65-75.
- SHEN Q, ZHANG L, ZHANG Y X, et al. Lightweight distracted driving behavior detection method based on improved YOLOv8n[J]. Electronic Measurement Technology, 2024, 47(24): 65-75.
- [100] 罗显志, 汪航. 跨尺度特征融合的无人机小目标检测算法[J/OL]. 计算机工程与应用[2025-04-05]. <https://kns.cnki>.

- net/kcms/detail/11.2127.tp.20250326.1510.018.html.
- LUO X Z, WANG H. Small target detection algorithm for UAV based on cross-scale feature fusion[J/OL]. Computer Engineering and Applications[2025-04-05]. <https://kns.cnki.net/kcms/detail/11.2127.tp.20250326.1510.018.html>.
- [101] 史丽晨, 杨超, 刘雪超, 等. 基于CDD-YOLO的轻量级低光照目标检测算法[J]. 计算机工程与应用, 2025, 61(6): 106-117.
- SHI L C, YANG C, LIU X C, et al. Lightweight low-light object detection algorithm based on CDD-YOLO[J]. Computer Engineering and Applications, 2025, 61(6): 106-117.
- [102] 钱尚乐, 曹伟, 高军伟. 基于改进YOLOv7-tiny的轻量化列车轮对踏面缺陷检测方法[J]. 光电子·激光, 2025, 36(7): 733-744.
- QIAN S L, CAO W, GAO J W. Lightweight train wheelset tread defect detection method based on improved YOLOv7-tiny[J]. Journal of Optoelectronics·Laser, 2025, 36(7): 733-744.
- [103] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017, 60(6): 84-90.
- [104] 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.
- [105] LI D, LI L, CHEN Z, et al. Shift-ConvNets: small convolutional kernel with large kernel effects[EB/OL]. [2025-02-15]. <https://arxiv.org/abs/2401.12736>.
- [106] HE X J, JIN J, JIANG Y, et al. A lightweight convolutional neural network-based feature extractor for visible images[J]. Computer Vision and Image Understanding, 2024, 249: 104157.
- [107] TANG A X, WANG Z Y, TIAN S G, et al. Series arc fault identification method based on lightweight convolutional neural network[J]. IEEE Access, 2024, 12: 5851-5863.
- [108] WANG H L, QIAN H M, FENG S, et al. L-SSD: lightweight SSD target detection based on depth-separable convolution [J]. Journal of Real-Time Image Processing, 2024, 21(2): 33.
- [109] ZHANG H, LAI S Q, WANG Y X, et al. SCGNet: shifting and cascaded group network[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2023, 33(9): 4997-5008.
- [110] 陈金荣, 许燕, 周建平, 等. 基于YOLO-SSAR的自然环境下红花检测算法[J]. 农业工程学报, 2025, 41(2): 215-223.
- CHEN J R, XU Y, ZHOU J P, et al. Detecting safflower in the natural environment using YOLO-SSAR[J]. Transactions of the Chinese Society of Agricultural Engineering, 2025, 41(2): 215-223.
- [111] LECUN Y, DENKER J, SOLLÀ S. Optimal brain damage [C]//Advances in Neural Information Processing Systems 2, 1989.
- [112] HASSIBI B, STORK D. Second order derivatives for network pruning: optimal brain surgeon[C]//Advances in Neural Information Processing Systems 5, 1992.
- [113] SHI X, DING J, HAO Z, et al. Towards energy efficient spiking neural networks: an unstructured pruning framework[C]//Proceedings of the 12th International Conference on Learning Representations, 2024.
- [114] LUO J H, WU J X, LIN W Y. ThiNet: a filter level pruning method for deep neural network compression[C]//Proceedings of the 2017 IEEE International Conference on Computer Vision. Piscataway: IEEE, 2017: 5068-5076.
- [115] WANG B, KINDRATENKO V. RL-Pruner: structured pruning using reinforcement learning for CNN compression and acceleration[EB/OL]. [2025-02-15]. <https://arxiv.org/abs/2411.06463>.
- [116] FANG G F, MA X Y, SONG M L, et al. DepGraph: towards any structural pruning[C]//Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2023: 16091-16101.
- [117] LIU Z, SUN M, ZHOU T, et al. Rethinking the value of network pruning[EB/OL]. [2025-02-15]. <https://arxiv.org/abs/1810.05270>.
- [118] 亢洁, 常琦, 王勃, 等. 融合DepGraph偏移正则化的绝缘子多缺陷检测轻量化算法[J/OL]. 计算机工程与应用 [2025-04-19]. <https://kns.cnki.net/kcms/detail/11.2127.tp.20250326.1741.031.html>.
- KANG J, CHANG Q, WANG Q, et al. a lightweight algorithm for multi-defect detection in insulators with integrated DepGraph offset regularization[J/OL]. Computer Engineering and Applications[2025-04-19]. <https://kns.cnki.net/kcms/detail/11.2127.tp.20250326.1741.031.html>.
- [119] HINTON G, VINYALS O, DEAN J. Distilling the knowledge in a neural network[EB/OL]. [2025-02-15]. <https://arxiv.org/abs/1503.02531>.
- [120] ZHAO B R, CUI Q, SONG R J, et al. Decoupled knowledge distillation[C]//Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2022: 11943-11952.
- [121] YANG C G, ZHOU H L, AN Z L, et al. Cross-image relational knowledge distillation for semantic segmentation[C]//Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2022: 12309-12318.

- [122] PARK W, KIM D, LU Y, et al. Relational knowledge distillation[C]//Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2019: 3962-3971.
- [123] 王改华, 李柯鸿, 龙潜, 等. 基于知识蒸馏的轻量化 Transformer 目标检测[J]. 系统仿真学报, 2024, 36(11): 2517-2527.
- WANG G H, LI K H, LONG Q, et al. Object detection of lightweight transformer based on knowledge distillation[J]. Journal of System Simulation, 2024, 36(11): 2517-2527.
- [124] LIU H, SIMONYAN K, YANG Y. DARTS: differentiable architecture search[EB/OL]. [2025-02-15]. <https://arxiv.org/abs/1806.09055>.
- [125] RADOSAVOVIC I, KOSARAJU R P, GIRSHICK R, et al. Designing network design spaces[C]//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2020: 10425-10433.
- [126] PHAM H, GUAN M, ZOPH B, et al. Efficient neural architecture search via parameters sharing[C]//Proceedings of the International Conference on Machine Learning, 2018: 4095-4104.
- [127] CI Y Z, LIN C, SUN M, et al. Evolving search space for neural architecture search[C]//Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2021: 6639-6649.
- [128] WANG L, FONSECA R, TIAN Y. Learning search space partition for black-box optimization using monte carlo tree search[C]//Advances in Neural Information Processing Systems 33, 2020: 19511-19522.
- [129] REAL E, MOORE S, SELLE A, et al. Large-scale evolution of image classifiers[C]//Proceedings of the International Conference on Machine Learning, 2017: 2902-2911.
- [130] ZOPH B, VASUDEVAN V, SHLENS J, et al. Learning transferable architectures for scalable image recognition [C]//Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 8697-8710.
- [131] XU Y, XIE L, ZHANG X, et al. PC-DARTS: partial channel connections for memory-efficient architecture search[EB/OL]. [2025-02-15]. <https://arxiv.org/abs/1907.05737>.
- [132] YANG H, ZHANG Y S, YIN C B, et al. Ultra-lightweight CNN design based on neural architecture search and knowledge distillation: a novel method to build the automatic recognition model of space target ISAR images[J]. Defence Technology, 2022, 18(6): 1073-1095.
- [133] BOUTROS F, SIEBKE P, KLEMT M, et al. PocketNet: extreme lightweight face recognition network using neural architecture search and multistep knowledge distillation[J]. IEEE Access, 2022, 10: 46823-46833.
- [134] LIU D C, YAMASAKI T, WANG Y, et al. Toward extremely lightweight distracted driver recognition with distillation-based neural architecture search and knowledge transfer[J]. IEEE Transactions on Intelligent Transportation Systems, 2023, 24(1): 764-777.
- [135] 杨军, 韩鹏飞. 采用神经网络架构搜索的高分辨率遥感影像目标检测[J]. 吉林大学学报(工学版), 2024, 54(9): 2646-2657.
- YANG J, HAN P F. Object detection of high-resolution remote sensing images by neural architecture search[J]. Journal of Jilin University (Engineering and Technology Edition), 2024, 54(9): 2646-2657.
- [136] SHIN J, SO J, PARK S, et al. NIPQ: noise proxy-based integrated pseudo-quantization[C]//Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2023: 3852-3861.
- [137] LIU J W, NIU L, YUAN Z H, et al. PD-quant: post-training quantization based on prediction difference metric[C]//Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2023: 24427-24437.
- [138] 陈俊杰, 陈哲宇, 郑子滨, 等. 基于FPGA的高能效纸板缺陷检测系统[J]. 计算机测量与控制, 2025, 33(1): 45-52.
- CHEN J J, CHEN Z Y, ZHENG Z B, et al. High-energy efficiency cardboard defect detection system based on FPGA [J]. Computer Measurement & Control, 2025, 33(1): 45-52.
- [139] YANG H R, TANG M X, WEN W, et al. Learning low-rank deep neural networks via singular vector orthogonality regularization and singular value sparsification[C]//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Piscataway: IEEE, 2020: 2899-2908.
- [140] MO D M, WONG W K, LAI Z H, et al. Weighted double-low-rank decomposition with application to fabric defect detection[J]. IEEE Transactions on Automation Science and Engineering, 2021, 18(3): 1170-1190.
- [141] YIN M, SUI Y, LIAO S Y, et al. Towards efficient tensor decomposition-based DNN model compression with optimization framework[C]//Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2021: 10669-10678.
- [142] AHMED W, HAJIMOLAHOSEINI H, WEN A, et al. Speeding up resnet architecture with layers targeted low rank

- decomposition[EB/OL]. [2025-02-15]. <https://arxiv.org/abs/2309.12412>.
- [143] 林德铝, 刘畅, 陈琦, 等. 基于低秩分解的YOLO轻量化目标检测模型[J]. 机车电传动, 2024(1): 138-144.
- LIN D L, LIU C, CHEN Q, et al. Low-rank decomposition based lightweight YOLO model for the object detection [J]. Electric Drive for Locomotives, 2024(1): 138-144.
- [144] YANG F C, HUANG L D, TAN X W, et al. FasterNet-SSD: a small object detection method based on SSD model[J]. Signal, Image and Video Processing, 2024, 18(1): 173-180.
- [145] 岳永恒, 宁睿厚. 基于轻量化CenterNet的智能车辆目标检测算法[J]. 华南理工大学学报(自然科学版), 2024, 52(8): 45-55.
- YUE Y H, NING R H. Intelligent vehicle object detection algorithm based on lightweight CenterNet[J]. Journal of South China University of Technology (Natural Science Edition), 2024, 52(8): 45-55.
- [146] 廖宁生, 曹天秀, 刘科言, 等. 复合特征与多尺度融合的小目标检测算法[J]. 计算机工程与应用, 2025, 61(3): 111-120.
- LIAO N S, CAO T X, LIU K Y, et al. Small target detection algorithm for UAV based on composite feature and multi-scale fusion[J]. Computer Engineering and Applications, 2025, 61(3): 111-120.
- [147] 林永洪, 郑建明, 李照. 基于UD-DETR轻量化网络小目标异物检测算法[J]. 信息化研究, 2025, 51(1): 71-78.
- LIN Y H, ZHENG J M, LI Z. UD-DETR: a lightweight network algorithm for the detection of small foreign objects [J]. Informatization Research, 2025, 51(1): 71-78.
- [148] 刘延芳, 余佳宇, 袁秋帆, 等. 无人机遥感图像实时小目标检测方法[J]. 航空学报, 2024, 45(14): 59-78.
- LIU Y F, SHE J Y, YUAN Q F, et al. Real-time small target detection networks for UAV remote sensing[J]. Acta Aeronautica et Astronautica Sinica, 2024, 45(14): 59-78.
- [149] 周赵轩, 曹岩. ZZX-YOLO: 改进YOLOv7-tiny的钢材缺陷检测算法[J/OL]. 计算机工程与应用[2025-04-05]. <https://kns.cnki.net/kcms/detail/11.2127.tp.20250320.1327.008.html>.
- ZHOU Z X, CAO Y. ZZX-YOLO: improved YOLOv7-tiny steel defect detection algorithm[J/OL]. Computer Engineering and Applications[2025-04-05]. <https://kns.cnki.net/kcms/detail/11.2127.tp.20250320.1327.008.html>.
- [150] 胡玮, 赵菊敏, 李灯熬. 基于改进YOLOv7-Tiny的轻量化激光器芯片缺陷检测算法[J]. 太原理工大学学报, 2025, 56(1): 137-147.
- HU W, ZHAO J M, LI D A. A lightweight laser chip defect detection algorithm based on improved YOLOv7-tiny[J]. Journal of Taiyuan University of Technology, 2025, 56(1): 137-147.
- [151] 梁礼明, 龙鹏威, 卢宝贺, 等. EHH-YOLOv8s: 一种轻量级的带钢表面缺陷检测算法[J/OL]. 北京航空航天大学学报[2025-07-18]. <https://doi.org/10.13700/j.bh.1001-5965.2024.0426>.
- LIANG L M, LONG P W, LU B H, et al. EHH-YOLOv8s: a lightweight algorithm for strip surface defect detection[J/OL]. Journal of Beijing University of Aeronautics and Astronautics[2025-07-18]. <https://doi.org/10.13700/j.bh.1001-5965.2024.0426>.
- [152] 王春梅, 刘欢. YOLOv8-VSC: 一种轻量级的带钢表面缺陷检测算法[J]. 计算机科学与探索, 2024, 18(1): 151-160.
- WANG C M, LIU H. YOLOv8-VSC: lightweight algorithm for strip surface defect detection[J]. Journal of Frontiers of Computer Science and Technology, 2024, 18(1): 151-160.
- [153] 刚帅, 刘培胜, 郭希旺. 改进基于YOLOv8n的轻量化钢材表面缺陷检测算法[J]. 电子测量技术, 2025, 48(3): 74-82.
- GANG S, LIU P S, GUO X W. Lightweight improved YOLOv8n model for steel defect detection features[J]. Electronic Measurement Technology, 2025, 48(3): 74-82.
- [154] 许晓阳, 高重阳. 改进YOLOv7-tiny的轻量级红外车辆目标检测算法[J]. 计算机工程与应用, 2024, 60(1): 74-83.
- XU X Y, GAO C Y. Improved YOLOv7-tiny lightweight infrared vehicle target detection algorithm[J]. Computer Engineering and Applications, 2024, 60(1): 74-83.
- [155] 罗向龙, 吕温馨, 石镇岳, 等. 改进YOLOv8n的轻量化交通标志检测算法[J]. 激光与光电子学进展, 2025, 62(12): 422-433.
- LUO X L, LÜ W X, SHI Z Y, et al. Improved lightweight traffic sign detection algorithm for YOLOv8n[J]. Laser & Optoelectronics Progress, 2025, 62(12): 422-433.
- [156] 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.
- LIU F, ZHONG Y F, QIU J W. Lightweight traffic sign recognition and detection algorithm based on improved YOLOv5s[J]. Laser & Optoelectronics Progress, 2024, 61(24): 102-114.
- [157] 王泽宇, 徐慧英, 朱信忠, 等. 基于YOLOv8改进的密集行人检测算法: MER-YOLO[J]. 计算机工程与科学, 2024, 46(6): 1050-1062.
- WANG Z Y, XU H Y, ZHU X Z, et al. An improved dense pedestrian detection algorithm based on YOLOv8: MER-

YOLO[J]. Computer Engineering & Science, 2024, 46(6): 1050-1062.

[158] CHE C, ZHENG H, HUANG Z, et al. Intelligent robotic control system based on computer vision technology[EB/OL]. [2025-02-15]. <https://arxiv.org/abs/2404.01116>.

[159] 熊其冰, 苗启广, 杨天, 等. 一种基于混合量子卷积神经网络的恶意代码检测方法[J]. 计算机科学, 2025, 52(3): 385-390.

XIONG Q B, MIAO Q G, YANG T, et al. Malicious code detection method based on hybrid quantum convolutional neural network[J]. Computer Science, 2025, 52(3): 385-390.

[160] 王东. 基于神经网络的人脸识别模型研究[J]. 科技创新与应用, 2024, 14(22): 5-8.

WANG D. Research on face recognition model based on neural network[J]. Technology Innovation and Application, 2024, 14(22): 5-8.

[161] 喇超, 李森, 张峰, 等. 高效 CNN 加速器设计[J/OL]. 计算机科学与探索[2025-02-27]. <https://kns.cnki.net/kcms/detail/11.5602.TP.20250226.0956.002.html>.

LA C, LI M, ZHANG F, et al. MSNAP: a high efficiencies CNN accelerator[J/OL]. Journal of Frontiers of Computer Science and Technology[2025-02-27]. <https://kns.cnki.net/kcms/detail/11.5602.TP.20250226.0956.002.html>.

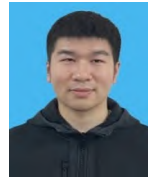
[162] SAMANTA A, HATAI I, MAL A K. A survey on hardware accelerator design of deep learning for edge devices[J]. Wireless Personal Communications, 2024, 137(3): 1715-1760.

[163] SUN K L, WANG X W, MIAO X, et al. A review of AI edge devices and lightweight CNN and LLM deployment[J]. Neurocomputing, 2025, 614: 128791.



董甲东(1968—),男,湖北宜昌人,博士,教授,主要研究方向为智能制造、工业互联网等。

DONG Jiadong, born in 1968, Ph.D., professor. His research interests include intelligent manufacturing, industrial Internet, etc.



桑飞虎(1998—),男,安徽阜阳人,硕士研究生,主要研究方向为目标检测、机器视觉等。

SANG Feihu, born in 1998, M.S. candidate. His research interests include target detection, machine vision, etc.



郭庆虎(1999—),男,安徽宿州人,硕士研究生,主要研究方向为深度学习、机器视觉等。

GUO Qinghu, born in 1999, M.S. candidate. His research interests include deep learning, machine vision, etc.



陈琳(1999—),女,安徽安庆人,硕士研究生,主要研究方向为工业互联网、边缘计算等。

CHEN Lin, born in 1999, M.S. candidate. Her research interests include industrial Internet, edge computing, etc.



郑春香(1973—),女,湖北宜昌人,硕士,副教授,主要研究方向为计算机应用。

ZHENG Chunxiang, born in 1973, M.S., associate professor. Her research interest is computer applications.