DOI: 10.13382/j.jemi.B2407857

高性能实时轻量化嵌入式缺陷检测网络的构建*

许志杰 吴黎明 张巧芬 王桂棠
(广东工业大学机电工程学院 广州 510006)

摘 要:针对工业嵌入式场景中缺陷检测模型存在参数量大、计算复杂度高与实时性要求之间的矛盾,提出由跨阶段部分卷积(CSPPC)模块、卷积跨尺度特征融合模块(CCFM)及 SA_Detect 融合模块构建 CCS-YOLO 轻量化缺陷检测网络,通过设计消融实验和对比实验验证其轻量化性能。为增强在处理复杂视觉任务时的特征提取与表达能力并结合部分卷积操作优化模型的性能与效率采用 CSPPC 模块,融合不同尺度的特征提升模型对尺度变化的适应性和对小尺度对象的检测能力采用 CCFM 模块,进一步减少模型参数量实现模型轻量化采用融合共享卷积的 SA_Detect 模块,有效提升特征表达、目标定位和分类性能。实验结果表明,CCS-YOLO 模型与 YOLOv8n 相比,模型大小、计算量和权重参数分别减少了 56.7%、51.9%和 54.0%,轻量化效果显著,并在 RK3568 嵌入式平台上部署检测速度维持在 37 fps 以上,实时性能得到验证,实用高效。可见系统的应用性价比得到提高,有效克服精度稍微下降带来的不足,而所构建的缺陷检测网络 CCS-YOLO 能够解决工业嵌入式场景中的资源受限问题,实现低算力设备达到高性能实时轻量化的可行方案,具有重要的工程价值。

关键词: 轻量化缺陷检测;部分卷积;特征融合;共享卷积;YOLOv8n
中图分类号: TP391.41;TN41 **文献标识码**: A **国家标准学科分类代码**: 520.6040

Construction of high-performance real-time lightweight embedded defect detection network

Xu Zhijie Wu Liming Zhang Qiaofen Wang Guitang
(School of Electromechanical Engineering, Guangdong University of Technology, Guangzhou 510006, China)

Abstract: Aiming at the contradiction between the large number of parameters, high computational complexity and real-time requirements of defect detection models in industrial embedded scenarios, a CCS-YOLO lightweight defect detection network is proposed to be constructed by CSPPC module, CCFM module and SA_Detect fusion module. Its lightweight performance is verified by designing ablation experiments and comparative experiments. In order to enhance the feature extraction and expression capabilities when processing complex visual tasks and combine partial convolution operations to optimize the performance and efficiency of the model, the CSPPC module is used. The CCFM module is used to fuse features of different scales to improve the model's adaptability to scale changes and the ability to detect small-scale objects. The SA_Detect module that fuses shared convolutions is used to further reduce the number of model parameters and achieve model lightweight, which effectively improves feature expression, target positioning and classification performance. The experimental results show that compared with YOLOv8n, the model size, computational complexity and weight parameters of the CCS-YOLO model are reduced by 56.7%, 51.9% and 54.0% respectively, with a significant lightweight effect. The detection speed is maintained above 34 fps when deployed on the RK3568 embedded platform, and the real-time performance is verified, which is practical and efficient. It can be seen that the application cost-effectiveness of the system has been improved, effectively overcoming the shortcomings caused by a slight decrease in accuracy. The constructed defect detection network CCS-YOLO can solve the problem of resource constraints in industrial embedded scenarios and realize a feasible solution for low-computing power devices to achieve high performance, real-time and lightweight, which has important engineering value.

收稿日期:2024-09-27 Received Date: 2024-09-27
*基金项目:广东省科技计划项目(2019B01001017)、2023 年佛山市促进高校科技成果服务产业发展扶持项目(2023DZXX02)、佛山市南海区 2024 年(第 14 批)创新创业人才团队项目资助

Keywords: lightweight defect detection; partial convolution; feature fusion; shared convolution; YOLOv8n

0 引 言

目标检测是计算机图像视觉领域中重要的研究方向,已经从早期的基于规则的算法演进到现代基于深度学习的方法。高计算资源需求、高能耗、以及部署和维护困难的问题,往往在工业场景应用中难于克服,而轻量化的缺陷检测模型以其高检测效率、低运算成本和易于部署的优势,逐渐成为主流选择。

工业产品的表面缺陷检测传统是依赖人工筛查,效率低下且出错率高。引入机器学习技术自动实现特征提取显著提升检测效率,深度学习的快速发展使之高效智能化。目前主流的目标检测算法分为基于回归的单阶段目标检测算法[1-3]和基于候选区域的两阶段目标检测算法[4]。虽然两阶段算法在精度上有优势,但单阶段算法的更快推理速度,适用于硬件资源受限的嵌入式缺陷实时检测。

近年来工业现场应用高精度高推理速度的缺陷检测模型成为研究热点,如 LF-YOLO 模型[5]设计局部填充上采样模块,通过结合插值和自校正卷积来优化特征图的质量,减少信息丢失提升小目标的检测能力,但该设计方法会增加计算复杂度,尤其是自校正卷积部分;轻量级 DCN-YOLO 模型[6]将可形变卷积网络(deformable convolution network, DCN)与 YOLOv5 结合引入可学习的偏移量来灵活调整卷积核的形状,自适应地处理不规则形状的目标,但 DCN 需要处理更多的计算任务和更复杂的特征映射,计算量和内存占用大幅增加。与 YOLOv8n 模型相比,计算成本高昂,不利于嵌入式平台的实时应用。

针对计算复杂度和参数量问题,MobileNetV3 作为特征提取主干被引入多个模型。例如 light-PDD 模型[7]和 SS-YOLO 模型[8]均采用 MobileNetV3,通过深度可分离卷积可降低计算复杂度和参数量,提升特征提取能力。light-PDD 模型设计双域注意机制有效解决微小尺寸 PCB 缺陷检测的难题,但也增加计算复杂度。SS-YOLO 模型引入 D-SimSPPF 模块和 SimAM 注意力机制,平衡空间和通道特征的提取,但也需要较大的计算资源。类似地,YOLO-SAGC 模型[9]结合了自注意力、图形卷积和深度可分离卷积,从而增加计算复杂度。尽管这些模型在精度和速度方面表现不错,但对于计算资源和存储空间有限的嵌入式设备而言,模型的计算复杂度和存储需求仍然较高。

相比之下,ESP-YOLO 模型[10]结合 YOLO、ELSAN、SE、PConv 和改进的 Soft_NMS,优化了骨干网络和颈部网络,使模型尺寸更小,且检测精度较好;Multi-CR 模型[11]通过引入 Multi-CR 块、SDDT-FPN 特征融合、PCR 模块和 C5ECA 模块,提高小目标缺陷的检测精度和特征融合能力,并实现轻量化;DCR-YOLO 模型[12]通过引入 CR 残差块、SDDT-FPN 特征融合和 PCR 模块,增强小目标缺陷的检测精度和自适应特征提取能力。上述模型在轻量化方面有所提升,但在资源受限的嵌入式设备上部署仍然存在较大困难。

在 YOLOv8 算法的最轻量级版本 n 的基础上,研究构建轻量化缺陷检测网络 CCS-YOLO,并在带有神经网络处理器的瑞芯微 RK3568 上实现嵌入式设备部署,提升计算效率和压缩能力,从而满足资源受限的嵌入式设备部署与应用需求。

1 CCS-YOLO 模型设计

广泛应用于视觉图像检测的 YOLO 作为一种端到端的深度学习目标检测模型,能够直接从输入图像中预测目标的类别概率和回归框。YOLOv8 的超轻量级 n 版本不仅具有较高检测精度还具备较高的实时性和轻量化的网络结构,其结构主要由主干网络(backbone)进行提取多层次特征表征,再利用特征融合层(neck)实现跨尺度的语义交互与空间增强,最终由检测头(head)基于优化后的特征图同步完成目标的分类置信度预测与边界框坐标回归。如图 1 所示,轻量化检测网络结构 CCS-YOLO 在设计上进行了多方面的优化,以提升模型在复杂视觉任务中的性能和效率。通过引入跨阶段部分卷积模块(cross stage partial convolution, CSPPC)增强特征提取和表达能力、提升模型在复杂视觉任务中的表现;从 Backbone 层到 Neck 层中间增加一个 Conv 统一通道数为 256,可降低网络深度和减少计算量,而基于卷积的跨尺度特征融合模块(convolution-based cross-scale feature fusion module, CCFM)融合不同尺度的特征能提高模型对尺度变化的适应性;进行目标定位和分类通过自注意力机制检测头(self-attention detect head, SA_Detect)实现。

1.1 CSPPC 模块设计

CCS-YOLO 网络的 Backbone 主干层由 CBS、CSPPC 和 SPPF 组成,CBS 负责提取特征并通过下采样操作逐步减小特征图的空间尺寸,更好捕捉不同尺度的特征信息;CSPPC 通过部分跨阶段连接和残差结构,有效减少计算量和参数规模,提升特征表达能力;SPPF 通过多尺度池化操作,增强网络的多尺度特征表达能力,应对不同大小的目标物体。

CSPPC 模块的设计融合多种深度学习技术,其核心

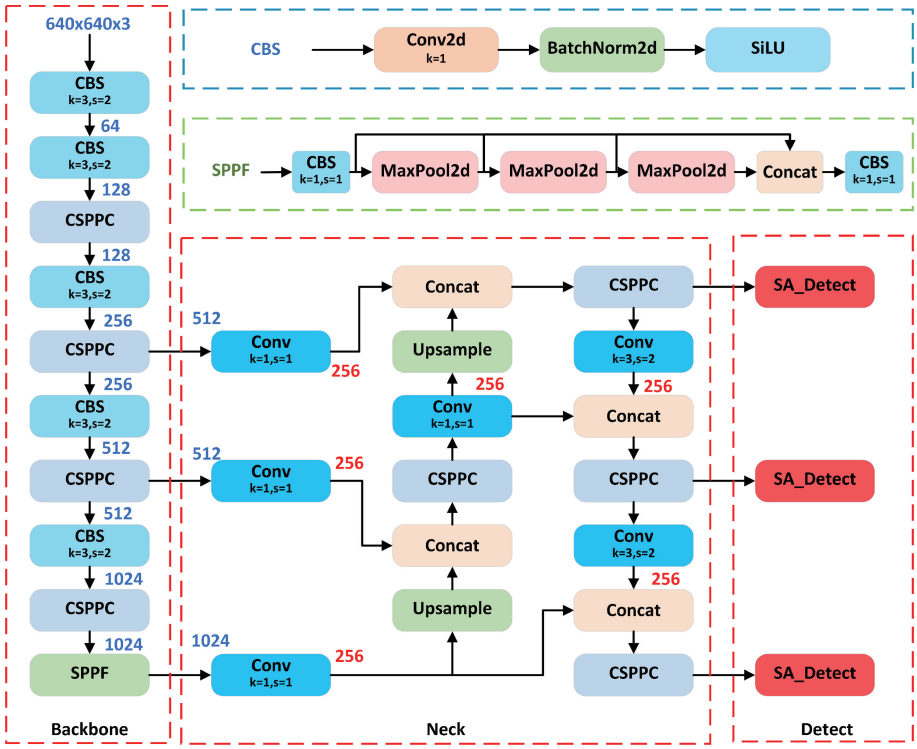


图 1　CCS-YOLO 网络结构

Fig. 1　CCS-YOLO network structure

部分包括部分卷积(partial convolution,PConv)和跨阶段部分连接(cross stage partial connections,CSP)。该模块基于 CSP 改进设计,结合 PConv、残差连接和跨阶段特征融合,有效提升神经网络的特征提取能力和计算效率,如图 2 所示。为了减少冗余计算,同时增强特征表示的能力,采用部分卷积和跨阶段连接的方法,先通过 1×1 的 Conv 调整输入特征图的通道数并将划分两份取其中一份进行两次 PConv,再将得到的特征图进行跨阶段特征融合。最后通过 1×1 的 Conv 调整输出通道数,降低模型的复杂性实现模型轻量化,在保持高精度的同时降低计算量和参数量,便于模型在嵌入式平台部署。

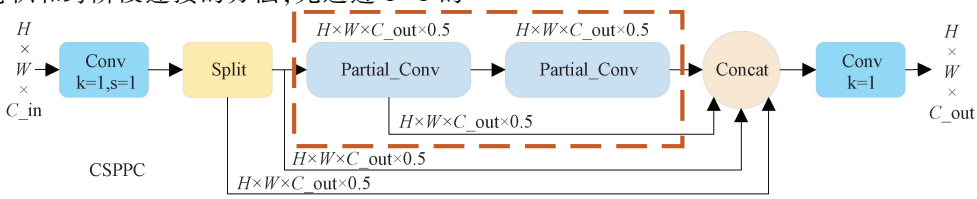


图 2　CSPPC 模块结构

Fig. 2　Structure diagram of CSPPC module

PConv 设计可有效提取空间特征,同时减少冗余的计算和内存访问,使缺陷检测模型更加轻量化。假设输入特征图 F_{in} 维度为 $H \times W \times C$,H 和 W 分别为特征图的高度和宽度,C 为通道数,PConv 结构如图 3 所示。PConv 将输入通道 C 分割成 4 份,取其中一份进行卷积计算再拼接输出特征图。在传统卷积中,所有通道都参与卷积计算,具有较高的计算复杂度和内存占用,为减少计算的卷积核数量,通过 PConv 对一部分通道进行卷积,剩余其他通道可以保留关键的特征信息。

在 CCS-YOLO 网络结构中,CSPPC 模块由两个标准 Conv、一个 Split 和两个 PConv 组成,能够高效融合特征并提升性能,适用于对计算资源受限的嵌入式平台,为缺陷检测模型的轻量化和高效部署提供支持。

1.2　CCFM 模块设计

CCFM 模块通过改进 Neck 特征融合层结构,将不同尺度的特征融合起来,增强模型对尺度变化的适应性和小尺度对象的检测能力。CCFM 模块结构如图 4 所示,根据特征金字塔网络(feature pyramid network,FPN)和路

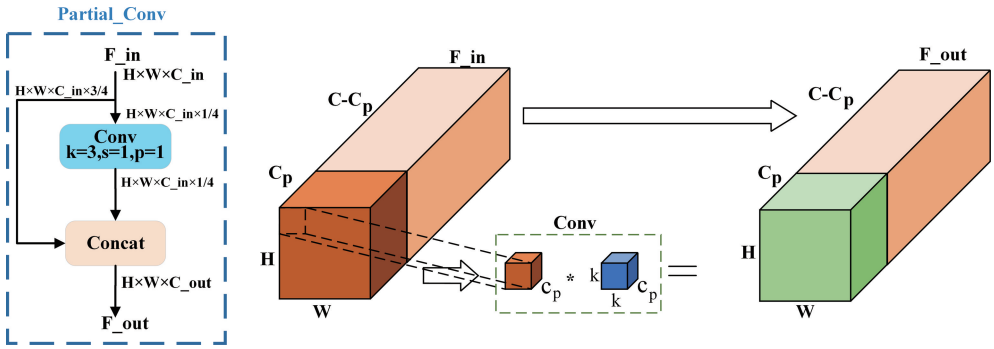


图 3　PConv 结构

Fig. 3　PConv structure

径聚合网络(path aggregation network,PAN)划分为 FPN_blocks 和 PAN_blocks,主要由 Conv 进行特征通道调整和特征投影,将调整后的信息进行特征融合,再由 CSPPC 模块提取融合后更具代表性的特征,增强不同特征图之间的信息流动,减少冗余信息,降低 Neck 特征融合层计算复杂度。FPN_Blocks 的主要功能包括将深层次特征图的通道数调整为 256,并通过上采样操作提升分辨率;利用 Conv 对 Backbone 层提取的特征进行投影,以保持多尺度特征图的一致性;将上采样后的特征与投影后的特征拼接,融合深层次语义信息和浅层细节信息;最后,通过 CSPPC 模块进一步处理,增强特征表达能力并降低计算量。PAN_Blocks 由横向连接 Conv 调整通道数为 256,再通过下采样卷积层降低分辨率,将降采样后的低分辨率特征图与通道数调整后的特征图拼接,经过 CSPPC 模块处理,增强多尺度信息的捕捉能力。CCFM 模块通过将特征图通道数统一为 256,在提升多尺度目标检测性能的同时实现模型轻量化,既保证特征融合的有效性,又维持缺陷检测的准确率。

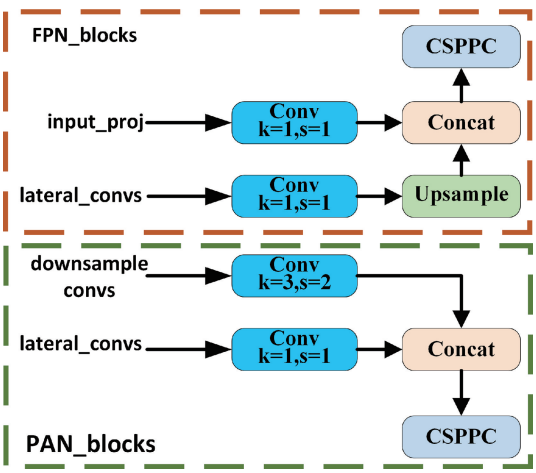


图 4　CCFM 模块结构

Fig. 4　Structure diagram of CCFM module

1.3　SA_Detect 模块设计

SA_Detect 模块是 CCS-YOLO 模型的 Head 层,主要对不同尺度的特征图进行目标检测,结合自注意力机制(self-attention,SA)和分布焦点损失(distribution focal loss, DFL)再通过特征提取和解码过程,输出目标的边界框和类别预测。SA_Detect 模块主要由共享 Conv 层、边界框回归分支和类别预测分支,SA_Detect 模块结构如图 5 所示。共享 Conv 层设计减少模型的冗余计算和共享特征表示,并提高模型的泛化能力,最后通过两个分支分别预测目标的边界框坐标和目标所属的类别。

SA 是 SA_Detect 模块的核心创新之一,主要用于增强特征图中的全局依赖关系。计算输入特征图中各个位置之间的关系(即注意力权重),使模型能够关注特征图中的重要区域。先通过输入特征向量与可学习的投影矩阵 $\boldsymbol{W}_Q$、$\boldsymbol{W}_K$、$\boldsymbol{W}_V$ 计算 $\boldsymbol{Q}$(query)、$\boldsymbol{K}$(key)、$\boldsymbol{V}$(value):

$$\begin{cases}\boldsymbol{Q}=\boldsymbol{W}_Q\cdot X\\ \boldsymbol{K}=\boldsymbol{W}_K\cdot X\\ \boldsymbol{V}=\boldsymbol{W}_V\cdot X\end{cases}\tag{1}$$

然后计算 $\boldsymbol{Q}$ 和 $\boldsymbol{K}$ 的内积,计算特征相关性,为了防止注意力得分过大,除以 Key 的向量维度 d_k 的平方根进行内积值缩放,再通过 Softmax 函数归一化,最后将 Softmax 矩阵与 V 相乘得到注意力得分矩阵为:

$$Attention(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\mathrm{Softmax}\left(\frac{\boldsymbol{Q}\cdot\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\cdot\boldsymbol{V}\tag{2}$$

最终的输出是输入特征图和注意力加权后的特征图的加权和为:

$$Output=\gamma\cdot Attention(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})+\boldsymbol{X}\tag{3}$$

式中:γ 是可学习的权重参数,用于调节注意力机制的输出与原始输入特征的比例。在输出中加入原始输入特征 X,通过残差连接的方式保留原始的局部信息。当 γ 值较小时,模型倾向于保留原始的局部特征;当 γ 值较大时,注意力机制的影响显著,模型会重视全局上下文信息。

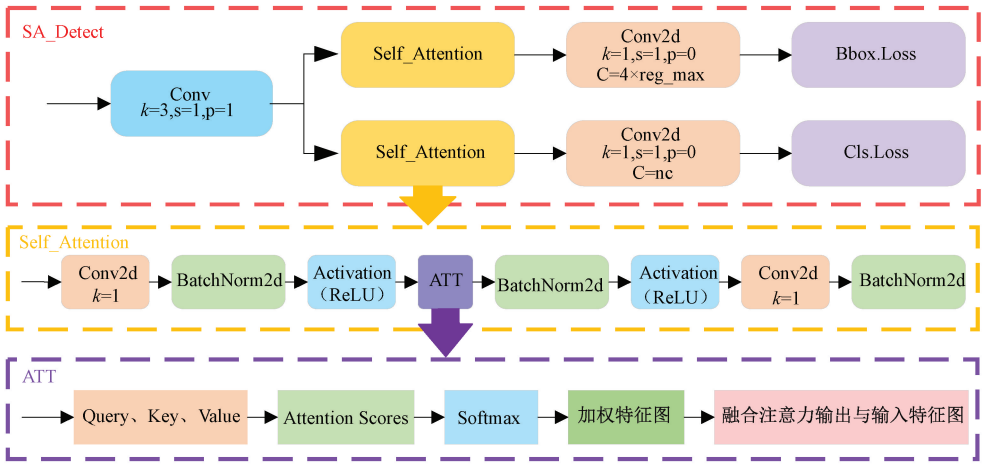


图 5　SA_Detect 模块结构

Fig. 5　Structure diagram of SA_detect module

SA 通过增强特征图的全局信息，使得 CCS-YOLO 模型在 Head 部分的特征提取更加高效和准确。减少对局部信息的过度依赖，提升模型在复杂场景中的检测性能，尤其是在多尺度特征融合方面，使得小目标的检测更加精确。

DFL 是用于提升边界框预测精度的关键技术。传统方法直接回归边界框的坐标，而 DFL 将每个坐标预测视为一个分布的积分结果。假设每个边界框的预测由分布 $p(x)$ 表示，预测值为：

$$\hat{x} = \sum_{i=1}^{N} i \cdot p(x = i) \tag{4}$$

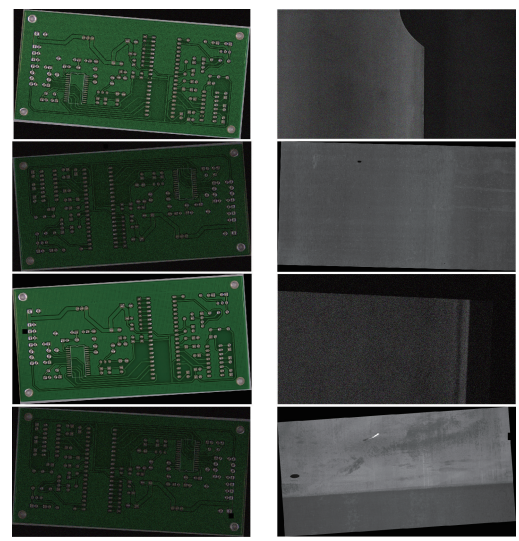
式中：$p(x=i)$ 表示预测值为 i 的概率。DFL 通过将边界框的预测视为一个分布的积分结果，而非直接回归边界框的坐标。

DFL 由预测值的概率分布来优化边界框预测的精度，并细化边界框的预测，减少由于直接回归带来的偏差，提升 CCS-YOLO 模型的 Head 在边界框回归方面的精度，对小目标和复杂目标的边界定位，有显著的效果提升。

2　实验结果与分析

2.1　数据集介绍

采用具有工业代表性的 PCB 数据集和 GC10-DET 数据集进行实验，可验证所构网络建模型的有效性。PCB 数据集[13]是一个公共合成数据集，包含 1 386 张图像，6 种缺陷，即缺失孔（MH）、鼠咬（MB）、开路（OC）、短路（Sh）、杂散（Sp）、杂铜（SC），取其中 693 张图像模拟实际工业环境中可能出现的光照变化、噪声干扰和几何变换，通过加噪声、调整亮度、旋转和裁剪等操作将数据集扩充到 11 008 张；GC10-DET[14]是在真实工业钢材中收集的表面缺陷数据集，包含 10 种缺陷类型，即冲孔（Pu）、焊缝线（Wl）、月牙弯（Cg）、水斑（Ws）、油斑（Os）、丝斑（Ss）、夹杂物（In）、轧坑（Rp）、折痕（Cr）和腰折（Wf）。为增强网络模型的泛化能力和鲁棒性，经人工筛选后对数据集进行加噪声、调整亮度、旋转和裁剪等操作扩充到 13 996 张，尺寸均为 2 048×1 000 的缺陷图片。将 PCB 数据集和 GC10-DET 数据集划分比例为 8∶1∶1 的训练集、验证集、测试集进行实验，图 6 所示为扩充数据集的部分示例图。



(a) PCB数据集　(b) GC10-DET数据集

(a) PCB datasets　(b) GC10-DET datasets

图 6　扩充数据集的部分示例

Fig. 6　Example plot of part of the expanded dataset

2.2　训练环境与嵌入式平台

实验模型训练基于 Visual Studio Code 的 SSH 功能远

程连接 Docker 容器进行，相关软件及其硬件配置如表 1 所示。在进行缺陷检测模型训练时，参数设置如下：输入图片尺寸为 640×640；epoch 为 300；学习率为 0.01；动量为 0.973；衰减系数为 0.000 5。

表 1 软件及硬件配置

Table 1 Software and hardware configuration

配置	版本
操作系统	Window11
解析器	Python 3.8.12
框架	Pytorch 1.11.0
容器	nvidia-deep-learning-container
加速环境	CUDA 11.5
内存	64 G
GPU	NVIDIA GeForce RTX 3080
CPU	Intel(R) Core(TM) i7-12700KF

模型部署的嵌入式设备是一款高性能低功耗可用于轻量级人工智能应用的国产瑞芯微芯片 RK3568，并内置 1Tops 算力的独立 NPU。相比于其他基于 ARM 内核的神经网络部署芯片，NVIDIA 公司的 Jetson Nano[15] 虽然拥有 TensorRT 推理加速框架，使神经网络模型的优化和部署简单，但价格较高，成本昂贵；树莓派的开发板价格较低，但只能使用 CPU 进行计算，且性能并不突出。RK3568 嵌入式平台在价格和性能取得较好的平衡，将 PC 端训练好的神经网络模型转换成 ONNX 模型格式，再使用相应工具链将 ONNX 模型进一步转换为可部署 RKNN 模型。

2.3 评价指标

用于全面评估所构建网络模型性能及泛化能力的指标包括精确度(precision，P)、召回率(recall，R)、平均检测精度均值(mean average precision，mAP)、模型计算量(GFLOPs)、模型参数量(Params)和模型权重大小(weight size)等。其中，精确度用于衡量模型预测的正类样本中，实际为正类的比例；召回率用来衡量所有实际正类样本中被正确预测为正类的比例；平均检测精度均值结合精确度和召回率，在不同的阈值下计算平均精度(average precision，AP)并取均值，衡量模型在不同类别上的检测性能；GFLOPs 表示模型每秒进行十亿次的浮点运算，衡量模型的计算复杂度，GFLOPs 越高，模型越复杂，需要计算资源越多；Params 衡量模型的复杂度和存储需求的重要指标，在资源受限的嵌入式平台，参数量对模型的实际应用有重要影响。性能评判指标相应计算公式如式(5)~(8)所示。

$$P = \frac{TP}{TP + FP} \tag{5}$$

$$R = \frac{TP}{TP + FN} \tag{6}$$

$$AP = \int_0^1 p(r)\,\mathrm{d}r \tag{7}$$

$$mAP = \frac{1}{N}\sum_{i=1}^{N} AP_i \tag{8}$$

式中：TP(true positives)表示真正例，即正确预测为正的样本数；FP(false positives)表示假正例，即错误预测为正的样本数；FN(false negatives)表示假负例，即错误预测为负的样本数；AP 是指精度召回率曲线下方的曲线的区域面积；N 为类别数；AP_i 为第 i 类的平均精度；mAP 指所有类别缺陷的平均 AP 值。

2.4 消融实验

通过逐步移除或替换关键组件设计消融实验，可观察 P、R、mAP、Params、计算量和 Weight size 的变化，评估各模块对模型性能的优化效果及统一通道数策略的有效性。实验主要从两个维度展开：首先，以 YOLOv8n 为基线模型，通过对比保留层级通道数和统一通道数两种配置下的模型性能，验证通道数统一策略的有效性，结果如表 2 所示；然后，基于基线模型，采用排列组合的实验设计方法，系统评估 CSPPC、CCFM 和 SA_Detect 3 个核心模块对算法性能的优化贡献度，结果如表 3 所示。

表 2 消融实验结果

Table 2 Results of ablation experiments

数据集	模型	P/%	R/%	mAP/%	Params/($\times 10^6$)	计算量/GFLOPs	Weight size/MB
PCB	保留层级通道数	94.5	87.5	93.3	3.00	8.1	6.3
	统一通道数(256)	93.2	88.2	92.9	2.00	6.6	4.2
	保留层级通道数(CCFM)	94.7	86.3	93.6	2.98	8.1	6.2
	统一通道数(256+CCFM)	94.1	86.7	92.7	1.96	6.6	4.2
GC10-DET	保留层级通道数	76.3	65.4	71.4	3.00	8.1	6.3
	统一通道数(256)	74.2	62.1	69.3	2.00	6.6	4.2
	保留层级通道数(CCFM)	77.1	64.1	71.5	2.98	8.1	6.2
	统一通道数(256+CCFM)	74.8	63.1	69.5	1.96	6.6	4.2

表 2 的消融实验可对比统一通道数与保留层级通道数的性能差异。在 PCB 数据集上的实验表明，将深层通道数统一为浅层一致通道数(256)后，模型的精度为 93.2%、召回率为 88.2%、mAP 为 92.9%，相较于原始保

表 3　消融实验结果

Table 3　Results of ablation experiments

数据集	模型			*P*/%	*R*/%	mAP/%	Params/(×10⁶)	计算量/GFLOPs	Weight size/MB
	CSPPC	CCFM	SA_Detect						
PCB				94.5	87.5	93.3	3.00	8.1	6.3
	√			93.7	85.6	91.2	2.12	5.9	4.5
		√		94.1	86.7	92.7	1.96	6.6	4.2
			√	96.2	90.3	93.8	2.86	7.0	6.1
	√	√		94.6	88.5	92.0	1.44	5.0	3.1
	√		√	93.9	87.2	91.6	1.98	4.8	4.2
		√	√	94.0	89.4	93.4	1.82	5.5	4.0
	√	√	√	94.7	87.3	93.2	1.30	3.9	2.9
GC10-DET				76.3	65.4	71.4	3.00	8.1	6.3
	√			74.0	62.5	68.9	2.12	5.9	4.5
		√		74.8	63.1	69.5	1.96	6.6	4.2
			√	75.5	63.9	70.2	2.86	7.0	6.1
	√	√		75.3	64.0	70.3	1.44	5.0	3.1
	√		√	75.9	64.5	70.7	1.98	4.8	4.2
		√	√	76.0	66.2	71.1	1.82	5.5	4.0
	√	√	√	76.7	65.1	71.0	1.30	3.9	2.9

留层级通道数模型(*P* 为 94.5%,*R* 为 87.5%,mAP 为 93.3%)略有下降。然而,模型参数、计算量和权重文件大小均得到显著提升,其中计算量降了约 20%,模型参数和权重文件大小均减少了 33%。这表明尽管统一通道数设计会降低浅层细节信息与深层语义信息的融合效率,从而对精度产生一定影响,但其能够大幅度减少模型的计算复杂度和内存占用。进一步实验显示,在原始模型中引入 CCFM 模块后,保留层级通道数的设计能够增强多尺度的特征融合能力,提供更强的适应性,使得模型精度和 mAP 略有提升,有效减缓多尺度信息融合效率下降的问题。最终,采用统一通道数(256)并结合 CCFM 模块的设计方案,其精度、召回率和 mAP 与原始模型的差距均控制在 1%以内,这一设计能够降低模型计算量和内存占用的同时保持较高的检测精度。在 GC10-DET 数据集上的实验表明,统一通道数策略虽然会削弱各层级间的语义信息融合能力,但通过引入 CCFM 模块设计,能够有效增强跨层次语义信息的交互与融合,从而提升模型精度。因此,采用统一通道数(256)并结合 CCFM 模块的设计方案,在模型性能与计算效率之间实现良好的平衡,适用于资源受限的嵌入式平台场景。

表 3 验证了 CCS-YOLO 模型中各个模块对轻量化和性能提升的贡献。实验结果表明,CSPPC、CCFM 和 SA_Detect 模块的引入及其组合,显著降低模型的参数量、计算量和权重大小,并保持较高的检测精度和召回率。具体而言,单独使用 CSPPC 模块能够有效减少冗余计算,降低参数量和计算量,但对精度有轻微影响;而 CCFM 模块进一步大幅度减少计算复杂度,但检测精度和召回率略有下降;SA_Detect 模块则通过增强特征表达能力,在维持高召回率的同时进一步优化了模型性能。最终,结合 3 个模块的 CCS-YOLO 模型在几乎保持原有精度、召回率和 mAP 的情况下,实现参数量减少 56.7%、计算量降低 51.9%以及权重大小缩减 54.0%的轻量化效果。这种改进使其更适用于工业场景中嵌入式设备部署,能够在有限的硬件资源下高效运行。

2.5　对比实验

为全面评估所构建网络结构的性能优势与创新价值,将 CCS-YOLO 模型与 YOLOv3[16]、YOLOv6[17]、YOLOv9c[18]等代表性网络进行对比分析,以验证其设计性能与先进性。为进一步量化模型在资源受限场景下的实际性能,引入自定义 mAP/计算量比值作为评价指标。该指标综合反映模型在单位计算量下的检测精度,能够更有效地衡量模型在计算效率与检测精度之间的平衡性。mAP/计算量比值越高,表明模型在维持高检测精度的同时具有更高的计算效率,从而更适合部署于计算资源受限的嵌入式或工业应用场景。

PCB 数据集的对比实验如表 4 所示,CCS-YOLO 模型以 mAP 为 93.2%检测精度优于 YOLOv3 和 YOLOv6,在轻量化方面表现突出,其模型参数量降至 1.30×10^6、计算量压缩至 3.9 GFLOPs,权重文件大小仅为 2.9 MB。尽管检测精度与 YOLOv9c 存在微弱差距(mAP = -0.9%),但其参数量与计算量较后者分别降低 94.9%和 96.2%,在低功耗边缘设备部署场景下实现效率-精度的最优平衡。此外,与 PPLCFaster-YOLOv5[19]和 YOLO-MBBi[20]相比,CCS-YOLO 在 mAP 上略有差距,但其 mAP/计算量比值(23.90)明显高于 PPLCFaster-YOLOv5 的 14.54 和 YOLO-MBBi 的 7.45,凸显其在计算资源受限

表 4　PCB 数据集对比实验结果

Table 4　Comparison experiment results for PCB dataset

模型	*P*/%	*R*/%	mAP/%	Params/(×10⁶)	计算量/GFLOPs	Weight size/MB	mAP/计算量
YOLOv3	93.5	88.3	92.1	12.1	18.9	24.4	4.87
YOLOv6	94.6	89.8	92.6	4.2	11.8	8.7	7.85
YOLOv9c	94.8	91.2	94.1	25.3	102.4	51.6	0.92
PPLCFaster-YOLOv5[19]	97.2	94.7	97.4	–	6.7	–	14.54
YOLO-MBBi[20]	95.8	94.6	95.3	–	12.8	–	7.45
CCS-YOLO(本文)	94.7	87.3	93.2	1.30	3.9	2.9	23.90

的环境下的应用潜力。

GC10-DET 数据集对比实验如表 5 所示,验证了 CCS-YOLO 的跨场景泛化能力,相较于 YOLOv3、YOLOv6 和 YOLOv9c 等基础架构,其参数量(1.30×10⁶)和计算量(3.9 GFLOPs)实现 2~3 个数量级的压缩突破。与工业级轻量模型 EML-YOLO[21] 相比,CCS-YOLO 不仅在检测精度上提升 4.4%,而且计算复杂度降低 3.1GFLOPs,实现效率与精度的双重突破。与 SRN-YOLO[22] 相比,CCS-YOLO 在仅牺牲 0.6% mAP 的前提下,以 1/41 参数量和 1/32 计算量实现单位算力精度(mAP/计算量)31.9 倍提升,验证了该架构在工业边缘部署场景中精度-效率均衡能力的优势。

表 5　GC10-DET 数据集对比实验结果

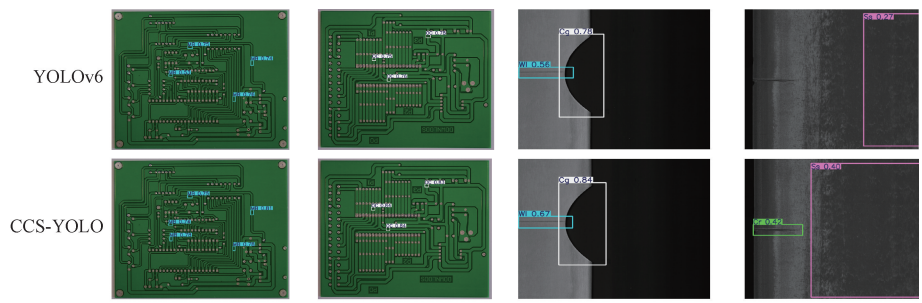
Table 5　Results of comparison experiments on GC10-DET dataset

模型	*P*/%	*R*/%	mAP/%	Params/(×10⁶)	计算量/GFLOPs	Weight size/MB	mAP/计算量
YOLOv3	75.1	63.3	69.8	12.1	18.9	24.4	3.69
YOLOv6	76.3	64.5	70.2	4.2	11.8	8.7	5.95
YOLOv9c	77.2	68.4	72.8	25.3	102.4	51.6	0.71
EML-YOLO[21]	–	–	66.6	2.7	7.0	–	9.51
SRN-YOLO[22]	–	–	71.6	54.0	126.3	–	0.57
CCS-YOLO(本文)	76.7	65.1	71.0	1.30	3.9	2.9	18.21

综合对比实验表明,CCS-YOLO 模型凭借 1.30×10⁶ 参数量与 3.9 GFLOPs 计算量的双维度压缩优势,以 18.2 高单位算力精度比值,展示出其高效的计算性能与精度-效率均衡能力。

2.6　可视化分析

为验证 CCS-YOLO 模型在工业缺陷检测中的有效性,选取 PCB 小目标(鼠咬、开路)及 GC10-DET 中多种复杂缺陷构建多维度测试集,与轻量化程度较高 YOLOv6 模型进行系统性对比,模型检测效果对比如图 7 所示。图 7(a)为鼠咬缺陷类型,YOLOv6 存在漏检的情况,而 CCS-YOLO 平均检测精度更高。图 7(b)为开路缺陷类型,两个模型没有出现漏检的情况,CCS-YOLO 模型对已检测出的缺陷在平均精度方面高于 YOLOv6 模型。图 7(c)为月牙弯和焊缝线缺陷类型,两个模型都没有出现漏检的情况,CCS-YOLO 模型对已检测出缺陷的平均精度方面高于 YOLOv6 模型。图 7(d)为折痕和丝斑缺陷类型,YOLOv6 存在漏检的情况,而 CCS-YOLO 模型能检测出折痕,且检测精度方面高。通过可视化对比分析,CCS-YOLO 模型具有更高检测效率和更少漏检率。



(a) 鼠咬　(b) 开路　(c) 月牙弯和焊缝线　(d) 折痕和丝斑

(a) Rat bite　(b) Open circuit　(c) Crescent bend and weld line　(d) Creases and silk spots

图 7　模型检测效果对比

Fig. 7　Comparison of model detection effect graph

2.7　嵌入式平台部署

将 CCS-YOLO 模型转换为 RKNN 模型,并量化为 int8 精度的模型部署到 RK3568 嵌入式平台上,验证模型的性能,RK3568 部署测试结果如表 6 所示。

表 6　RK3568 部署测试结果

Table 6　RK3568 deployment test results

数据集	模型	*P*/%	*R*/%	mAP/%	帧率/fps
PCB	YOLOv8n	92.7	86.3	92.3	35.6
	CCS-YOLO	93.5	86.8	92.8	38.2
GC10-DET	YOLOv8n	74.8	63.7	70.1	35.3
	CCS-YOLO	75.5	64.4	70.3	37.8

将 YOLOv8n 和 CCS-YOLO 模型量化为 int8 的 RKNN 模型后,检测性能比 PC 端略有下降,但 CCS-YOLO 的检测精度和 FPS 实时性都高。

3　结　论

为满足缺陷检测算法在工业应用中实时性、嵌入式设备部署需求,提出了一种针对工业场景应用的轻量化缺陷检测网络 CCS-YOLO。CSPPC 模块的引入能够融合部分卷积可有效减少模型的参数量和计算量,增强特征提取和表达能力;CCFM 模块的构建提高了模型对尺度变化的适应性,采用统一通道数策略降低了计算复杂度;采用共享卷积层并结合自注意力机制优化检测头结构,同时利用局部细节和全局上下文处理,有效提升了检测头的目标定位和分类性能。对 CSPPC 模块、CCFM 模块和 SA_Detect 模块进行了消融实验,实验验证了这些模块的有效性,与 YOLOv6、YOLOv9c 等经典模型及最新研究成果进行对比,实验结果表明,CCS-YOLO 模型参数量降至 1.30×10^6、计算量压缩至 3.9 GFLOPs,权重文件大小仅为 2.9 MB,轻量化效果显著。CCS-YOLO 在 PCB 和 GC10-DET 数据集上验证了跨场景泛化性能,其在 RK3568 嵌入式平台上检测速度可达 37 fps 以上,实现了模型在资源受限嵌入式设备上部署应用。构建的 CCS-YOLO 缺陷检测网络克服了轻量化带来的精度严重下降问题,实用性强性价比高,可应对嵌入式场景的算力资源挑战。进一步的研究将基于 FPGA 异构计算架构,构建算子级动态自适应机制,实现复杂工业场景下的超实时在线检测与边缘端自主决策闭环。

参考文献

[1] 赵佰亭, 张晨, 贾晓芬. ECC-YOLO:一种改进的钢材表面缺陷检测方法[J]. 电子测量与仪器学报, 2024, 38(4): 108-116.
ZHAO B T, ZHANG CH, JIA X F. ECC-YOLO: An improved method for detecting surface defects in steel[J]. Journal of Electronic Measurement and Instrumentation, 2024, 38(4): 108-116.

[2] 邝先验,程福军,吴翠琴,等. 基于改进 YOLOv7-tiny 的高效轻量遥感图像目标检测方法[J]. 电子测量与仪器学报,2024,38(7):22-33.
KUANG X Y, CHENG F J, WU C Q, et al. Efficient and lightweight target detection method for remote sensing images based on improved YOLOv7-tiny[J]. Journal of Electronic Measurement and Instrumentation, 2024, 38(7):22-33.

[3] HUSSAIN M. YOLOv1 to v8: Unveiling each variant-A comprehensive review of YOLO[J]. IEEE Access, 2024, 12: 42816-42833.

[4] SUN P, ZHANG R, JIANG Y, et al. Sparse R-CNN: An end-to-end framework for object detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(12): 15650-15664.

[5] 马肖瑶, 黎睿, 李自力, 等. 面向工业场景带钢表面缺陷检测的 LF-YOLO[J]. 计算机工程与应用, 2024, 60(18):78-87.
MA X Y, LI R, LI Z L, et al. LF-YOLO for strip surface defect detection in industrial scenes[J]. Computer Engineering and Applications, 2024, 60(18): 78-87.

[6] 卢俊哲, 张铖怡, 刘世鹏, 等. 面向复杂环境中带钢表面缺陷检测的轻量级 DCN-YOLO[J]. 计算机工程与应用, 2023, 59(15): 318-328.
LU J ZH, ZHANG CH Y, LIU SH P, et al. Lightweight DCN-YOLO for strip surface defect detection in complex environments[J]. Computer Engineering and Applications, 2023, 59(15): 318-328.

[7] TANG J, WANG Z, ZHANG H, et al. A lightweight surface defect detection framework combined with dual-domain attention mechanism[J]. Expert Systems with Applications, 2024, 238: 121726.

[8] LU J, YU M, LIU J. Lightweight strip steel defect detection algorithm based on improved YOLOv7[J]. Scientific Reports, 2024, 14(1): 13267.

[9] WANG G Q, ZHANG C Z, CHEN M S, et al. A high-accuracy and lightweight detector based on a graph convolution network for strip surface defect detection[J]. Advanced Engineering Informatics, 2024, 59: 102280.

[10] CHEN J, CHEN H, XU F, et al. Real-time detection of mature table grapes using ESP-YOLO network on embedded platforms[J]. Biosystems Engineering, 2024, 246: 122-134.

[11] 姜媛媛, 蔡梦南. 轻量化的印刷电路板缺陷检测网络 Multi-CR YOLO[J]. 电子测量与仪器学报, 2023,

37(11): 217-224.
JIANG Y Y, CAI M N. Lightweight PCB defect detection network Multi-CR YOLO [J]. Journal of Electronic Measurement and Instrumentation, 2023, 37(11): 217-224.
[12] JIANG Y, CAI M, ZHANG D. Lightweight network DCR-YOLO for surface defect detection on printed circuit boards[J]. Sensors, 2023, 23(17): 7310.
[13] DING R, DAI L, LI G, et al. TDD-Net: A tiny defect detection network for printed circuit boards[J]. CAAI Transactions on Intelligence Technology, 2019, 4(2): 110-116.
[14] LYU X, DUAN F, JIANG J JIA, et al. Deep metallic surface defect detection: The new benchmark and detection network[J]. Sensors, 2020, 20(6): 1562.
[15] QIAO W, GUO H, HUANG E, et al. Real-time detection of slug flow in subsea pipelines by embedding a YOLO object detection algorithm into jetson nano[J]. Journal of Marine Science and Engineering, 2023, 11(9): 1658.
[16] REDMON J, FARHADI A. YOLOv3: An incremental improvement[J]. ArXiv e-prints, 2018, DOI: 10.48550/arXiv.1804.02767.
[17] 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.
[18] CHIEN C T, JU R Y, CHOU K Y, et al. YOLOv9 for fracture detection in pediatric wrist trauma x-ray images[J]. ArXiv preprint arXiv: 2403.11249, 2024.
[19] 季堂煜, 赵倩, 赵琰, 等. 基于 PPLCFaster-YOLOv5 的 PCB 表面缺陷快检模型[J]. 电子测量技术, 2023, 46(11): 115-122.
JI T Y, ZHAO Q, ZHAO Y, et al. Rapid inspection model of PCB surface defects based on PPLCFaster-YOLOv5 [J]. Electronic Measurement Technology, 2023, 46(11): 115-122.
[20] DU B, WAN F, LEI G, et al. YOLO-MBBi: PCB surface defect detection method based on enhanced YOLOv5[J]. Electronics, 2023, 12(13): 2821.
[21] 苏佳, 贾泽, 秦一畅, 等. 面向工业表面缺陷检测的改进 YOLOv8 算法[J]. 计算机工程与应用, 2024, 60(14): 187-196.
SU J, JIA Z, QIN Y CH, et al. Improved YOLOv8 algorithm for industrial surface defect detection [J] Computer Engineering and Applications, 2024, 60(14): 187-196.
[22] GAO S, CHU M, ZHANG L. A detection network for small defects of steel surface based on YOLOv7[J]. Digital Signal Processing, 2024, 149: 104484.

作者简介

许志杰,2022 年于广东工业大学获得学士学位,现为广东工业大学硕士研究生,主要研究方向为智能感知、嵌入式系统。
E-mail: 1422411797@qq.com

Xu Zhijie received his B. Sc. degree from Guangdong University of Technology in 2022. Now he is a M. Sc. candidate at Guangdong University of Technology. His main research interests include intelligent sensing and embedded systems.

吴黎明(通信作者),2004 年于华南理工大学获得硕士学位,现为广东工业大学教授,主要研究方向为智能感知、嵌入式系统。
E-mail: jkyjs@gdut.edu.cn

Wu Liming (Corresponding author) received his M. Sc. degree from South China University of Technology in 2004. Now he is a professor at Guangdong University of Technology. His main research interests include intelligent sensing and embedded systems.