DOI: 10.13382/j.jemi.B2508346

基于 STR-DETR 的轻量化 PCB 缺陷检测算法*

陈枫赟[1,2]　李　鹏[1,2]

(1. 南京信息工程大学电子与信息工程学院　南京　210044;2. 南京信息工程大学
江苏省气象探测与信息处理重点实验室　南京　210044)

摘　要:针对现有印刷电路板缺陷检测模型参数量庞大、计算复杂度高,难以部署在计算资源有限的工业边缘设备上的问题,提出了一种基于 STR-DETR 的轻量化缺陷检测算法。首先,通过分组卷积重构轻量级网络 StarNet 形成新型主干网络 G-StarNet,在保留多尺度特征提取能力的同时显著减少模型复杂度;其次,在自适应特征交互模块中引入基于统计学特征的自注意力机制来代替原有的多头自注意力,降低了计算开销;再次,结合曼哈顿自注意力机制及其分解形式设计 RetBlockC3 模块,采用距离相关的衰减模式增强了局部特征的表达优先级,实现了计算复杂度从二次方到线性的优化;最后,提出了一种新的损失函数 FSN Loss,通过改善形状和尺度因素、样本分布不均对边界框回归结果的影响来增强检测的定位与分类准确性。实验结果表明,改进后的算法平均精度均值(mAP)mAP@0.5 达到了 96.7%,相较于基准模型,参数量减少了 50.8%,计算量下降了 55.4%,检测速度提高了 23.7%,验证了算法的有效性,能够满足轻量化小目标检测的需求。

关键词: RT-DETR;小目标检测;印刷电路板;轻量化;注意力机制

中图分类号: TP391;TN912　　**文献标识码**: A　　**国家标准学科分类代码**: 520.20

Lightweight PCB defect detection algorithm based on STR-DETR

Chen Fengyun[1,2]　Li Peng[1,2]

(1. School of Electronics and Information Engineering, Nanjing University of Information Science and Technology, Nanjing 210044, China;2. Jiangsu Key Laboratory of Meteorological Observation and Information Processing, Nanjing University of Information Science and Technology, Nanjing 210044, China)

Abstract: To address the challenges of existing PCB defect detection models, which suffer from excessive parameters, high computational complexity, and limited deploy ability on industrial edge devices with constrained computing resources, we propose a lightweight defect detection algorithm based on STR-DETR. First, we construct a novel backbone network, G-StarNet, by integrating group convolution into the lightweight StarNet architecture. This modification significantly reduces model complexity while preserving multi-scale feature extraction capabilities. Second, within the adaptive feature interaction module, a statistical feature-based self-attention mechanism replaces the conventional multi-head self-attention, effectively lowering computational overhead. Third, the RetBlockC3 module is designed by combining the Manhattan self-attention mechanism and its decomposed form. By incorporating a distance-dependent attenuation strategy, this module prioritizes local feature representation and reduces computational complexity from quadratic to linear scaling. Finally, we introduce a new loss function, FSN Loss, which mitigates the adverse effects of shape/scale variations and imbalanced sample distributions on bounding box regression, thereby enhancing both localization and classification accuracy. Experimental results demonstrate that the improved algorithm achieves an mAP@0.5 of 96.7%. Compared with the baseline model, it reduces parameters by 50.8%, computational load by 55.4%, and increases detection speed by 23.7%. These findings validate the algorithm's effectiveness in meeting the requirements of lightweight small-target detection tasks.

Keywords: RT-DETR; small target detection; printed circuit board; lightweight; attention mechanism

收稿日期: 2025-04-28　　Received Date: 2025-04-28

*基金项目:国家自然科学基金(41075115)、江苏省重点研发计划社会发展项目(BE2015692)、无锡市社会发展科技示范工程项目(N20191008)资助

0　引　言

印刷电路板(printed circuit board, PCB)作为电子工业中的核心组件,其设计与制造质量直接决定了电子设备的性能和可靠性。随着技术的不断进步,PCB 的设计与生产流程日益复杂,任何一个环节的工艺偏差都可能引发短路、断路、毛刺等微小缺陷[1]。这些缺陷的检测难度显著提升,而传统的人工检测方式存在效率低下、主观性强且容易误判的缺陷。因此,开发高效、精准的缺陷检测技术对于提升产品质量和优化制造成本具有重要的现实意义。

在现代工业制造中,PCB 缺陷检测主要分为传统检测和基于深度学习的检测方法。传统的 PCB 缺陷检测方法主要包括人工目视检查、电气性能测试[2]、超声扫描以及热成像检测等。人工目视检查依赖于操作人员的技能和视觉能力,不仅效率较低,而且成本较高。电气性能测试需要与元件建立物理电气连接,存在损坏元件的风险,且检测成本较高,难以满足高精度检测的需求。超声扫描和热成像检测虽然在某些方面具有一定优势,但设备成本高、检测速度慢,对表面缺陷的敏感度不足,且数据解析需要专业人员,难以适应小规模生产线的需求。

传统检测方法存在一定局限性,而基于深度学习的目标检测算法凭借其卓越的特征学习能力,在很大程度上弥补了这些不足,进而被广泛应用于各个领域。随着人工智能技术的不断进步,目标检测领域也在持续革新。基于深度学习的目标检测算法主要可划分为两阶段(two-stage)、单阶段(one-stage)和端到端(end-to-end)这 3 种类型。两阶段算法如 R-CNN[3]、Fast R-CNN[4]、Faster R-CNN[5]等,此类算法先产生候选框再识别物体,完成区域分类与边框回归,将检测精度置于优先地位,但因计算量大导致运行速度慢。陈仁祥等[6]基于 Faster R-CNN 展开研究,引入注意力机制以提升检测效果,在特征提取阶段采用分离注意力网络聚焦缺陷特征并减少噪声干扰,通过平衡特征金字塔融合不同分辨率特征,并利用非局部注意力机制增强缺陷表征、抑制噪声,模型的检测速度出现大幅度下降。单阶段算法例如 YOLO(you only look once)系列[7-10]、SSD[11]等,因密集预测缺乏精细化处理而导致精度不足。张莹等[12]提出的 YOLOPCB 网络,通过精简 YOLOv7 的主干网络结构,设计了跨通道信息连接模块和浅层特征融合模块,并采用自适应加权跳层连接策略来简化网络结构,这些改进虽然简化了网络结构,但也导致了信息丢失,使得检测精度有所下降。王军等[13]在 YOLOv8 的 C2f 模块中融入了 SE 注意力机制,以增强通道特征的权重,并在 SPPF 模块中引入了 Basic RFB 来扩大感受野,这些优化措施虽然提升了模型的性能,但在轻量化部署场景下,却带来了过高的资源消耗。传统两阶段与单阶段算法因依赖非极大值抑制(NMS)导致冗余框处理效率低、并行性差且阈值敏感易漏检,而端到端算法如 DETR[14]通过 Transformer 全局建模和消除 NMS 后处理,在简化流程的同时实现更高检测一致性。吴斌斌等[15]采用 CSPDarknet 替换原始骨干网络以增强特征提取能力,基于 CCFM 模块构建小目标增强金字塔结构 SOEP,通过增强的特征层和精细的特征融合提高模型对 PCB 小目标缺陷的定位与识别精度,但模型的参数量较大。

尽管现有方法在特定任务中展现出较高的模型精度,但其计算复杂度大幅增长,推理延迟显著增加,严重制约了其在边缘计算设备中的部署可行性,因此本文提出了一种轻量化的 STR-DETR 模型。

1　RT-DETR 模型

RT-DETR[16]是百度飞桨团队于 2023 年提出的基于 Transformer 架构的实时端到端目标检测模型,其模型结构如图 1 所示,由主干网络、高效混合编码器和带辅助预测头的 Transformer 解码器组成。

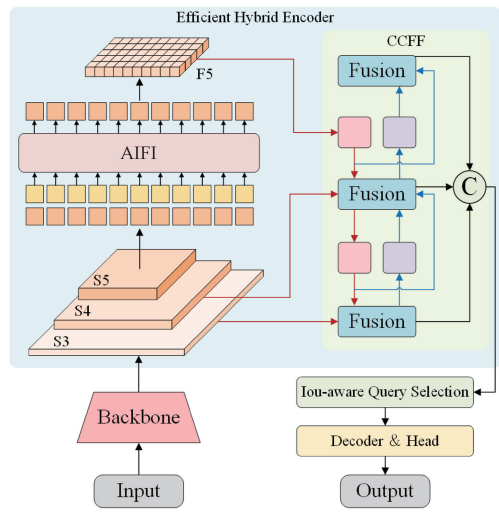


图 1　RT-DETR 模型结构

Fig. 1　RT-DETR model structure

该模型的主干网络采用卷积神经网络(convolutional neural network,CNN)架构,通过对输入图像的多阶段特征提取生成 S3、S4、S5 三层特征,这些特征图分别具有不同的空间分辨率和语义信息,为后续的混合编码器提供丰富的多尺度特征输入,以支持高效的目标检测任务。

高效混合编码器通过两大核心模块实现了计算效率与特征表达的平衡,基于注意力的尺度内特征交互模块(attention-based intra-scale feature interaction, AIFI)负

责对主干网络输出的最高层特征 S5 进行编码,通过单尺度 Transformer 编码器进行特征交互,在降低计算复杂度的同时精准捕捉高层语义信息;基于 CNN 的跨尺度特征融合模块(CNN-based cross-scale feature fusion, CCFF)则通过包含 1×1 卷积层(通道数调整)和 RepConv(特征融合)的多级融合块,以逐元素相加的方式将 S3、S4、S5 的跨尺度信息整合为全局特征序列,显著增强了模型对多尺度目标的表征能力。

解码阶段,模型通过交并比(IoU)感知查询选择机制从编码器输出中筛选出高置信度的初始对象查询,并引入不确定性最小化策略优化查询质量,确保初始特征的代表性。随后,借助多层 Transformer 的迭代优化,结合辅助预测头的中间监督,最终生成精确的目标边界框坐标与置信度分数。

RT-DETR 依据主干网络的不同可以划分为 ResNet 系列和 HGNetv2 系列,其中 ResNet 凭借残差连接与模块化设计有效地解决了深层网络的梯度退化问题,并根据网络层数细分为 R18、R34、R50、R101 等版本。针对工业场景中算法需满足低算力部署、高实时性及跨平台适配的核心需求,本文选择 RT-DETR-R18 作为基线模型。

2 STR-DETR 模型

尽管 RT-DETR 基准模型在实时目标检测任务中表现优异,然而其在计算资源受限的边缘设备(如嵌入式工控机、移动端处理器)中的部署仍面临计算量庞大的挑战,实际部署难度较大。针对这一问题,本文在深入分析 RT-DETR 结构的基础上,提出了轻量化的改进模型并命名为 STR-DETR,其网络结构如图 2 所示。首先,将分组卷积引入轻量化的 StarNet[17] 构成一种改进的神经网络 G-StarNet,替换原有的主干网络;其次,替换 AIFI 中的多头自注意力为一种基于统计学特征的自注意力机制;随后,设计结合了曼哈顿自注意力机制及其分解形式的 RetBlock 替代原模型 RepC3 模块中的 RepConv;最后,提出了一种新的损失函数 FSN(focaler-shape-NWD) Loss 来增强目标检测的定位与分类准确性。

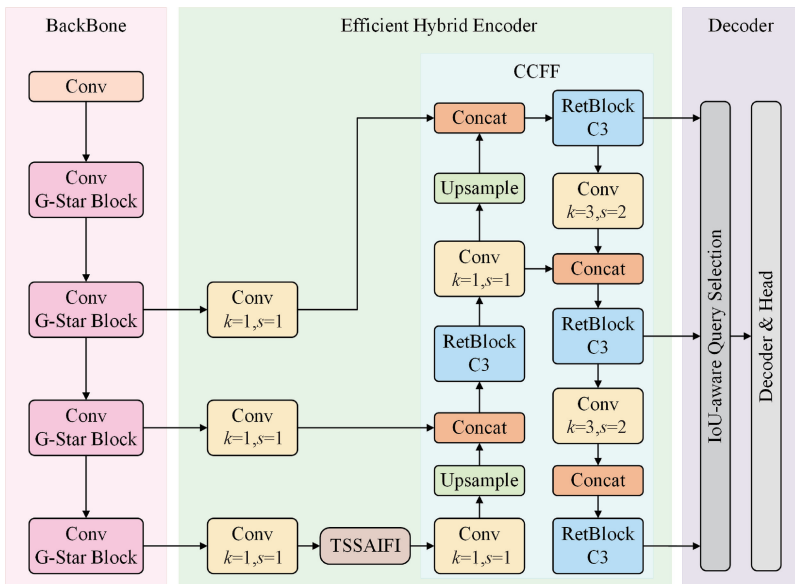


图 2 STR-DETR 模型结构

Fig. 2 STR-DETR model structure

2.1 G-StarNet 轻量化主干网络

StarNet 是一种基于“星操作”(star operation),即逐元素乘法构建的轻量化神经网络,其核心设计采用四阶段分层架构,无需复杂模块堆砌或精细超参数调整。星操作通过类似核技巧的机制,能在不扩展网络宽度的前提下将输入映射至高维非线性特征空间。这种特性使网络在低维空间计算时仍能捕捉高维特征关联性,为轻量化模型保持精度提供了新路径。该网络中使用的深度可分离卷积(depthwise separable convolution,DWConv)限制了通道间的特征融合和全局信息捕捉,因此将其用分组卷积(group convolution,GConv)代替,通过跨通道信息融合能力和更灵活的计算结构,提升了网络的特征提取能力和计算效率。G-StarNet 网络架构如图 3 所示。

单层中“星操作”定义为 $(\boldsymbol{W}_1^{\mathrm{T}}\boldsymbol{X}+B_1)*(\boldsymbol{W}_2^{\mathrm{T}}\boldsymbol{X}+B_2)$,该操作通过逐元素乘法,将输入特征从 d 维空间隐式映射至约 $\frac{(d+2)(d+1)}{2}\approx\left(\frac{d}{\sqrt{2}}\right)^2$ 维特征空间且满足 $d\geqslant 2$。其核心在于除部分特殊项外,输出中的每个特

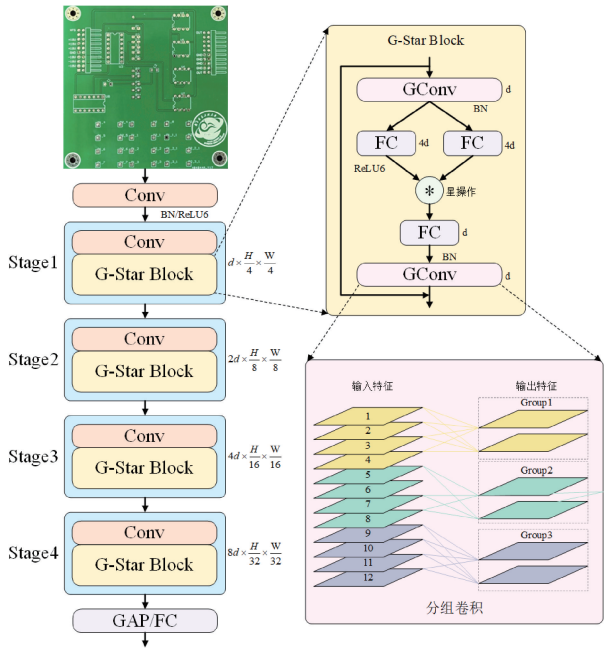


图 3　G-StarNet 网络架构

Fig. 3　G-StarNet network architecture

征项均与输入呈现非线性关联,从而在不显式增加网络宽度的前提下实现高维非线性表征。当堆叠多层星操作时,假设初始网络宽度为 d,经过 i 层后,隐式特征空间的维度可扩展至 $\boldsymbol{R}^{\left(\frac{d}{\sqrt{2}}\right)^{2^l}}$,例如一个 4 层、单层宽度为 128 的各向同性网络,通过星操作获得的隐式特征维度为 4.5×10^{15},层级堆叠后高效逼近超高维特征交互。相较于主流高效网络设计方法(如深度可分离卷积、特征复用等),星操作展现出独特优势:传统方法多通过减少通道数或简化卷积操作实现轻量化,但可能伴随特征表达能力下降。而星操作通过数学映射突破维度限制,在保持计算效率的同时实现高维特征交互。这种“低维计算—高维表达”的协同机制,为构建更紧凑高效的模型架构提供了新的理论切入点,推动轻量化网络设计突破现有技术框架的局限性。

分组卷积通过将输入和输出通道均划分为 G 组,仅允许同组通道交互,将参数量与计算量降至普通卷积的 $\frac{1}{G}$,其核心在于将输入特征图分组后,每组使用独立的卷积核进行稀疏连接(而非全通道密集连接)。假设输入的特征图尺寸为 $C\cdot H\cdot W$,输出特征图的数量为 N 个,则每个卷积核的通道数从 C 缩减为 $\frac{C}{G}$,总参数量由 $N\cdot C\cdot K^2$ 减少为 $\frac{N\cdot C\cdot K^2}{G}$。这种结构化稀疏机制不仅降低冗余,还能通过限制参数自由度增强模型正则化效果,从而在提升效率的同时维持甚至提升性能。

2.2　TSSAIFI 模块

原模型中 AIFI 模块使用的多头自注意力(multi-head self-Attention,MHSA)需要计算所有 Token 之间的两两相似性,这一过程虽然有效,但其资源开销显著,计算量和内存需求随着序列长度的增加而显著增加,产生较高的延迟导致边缘设备部署困难。为使模型的轻量化满足实时检测需求,引入了一种基于统计学特征的自注意力机制(token statistics self-attention,TSSA)[18],作为 ToST(token statistics transformer)的核心模块,其仅依靠 Token 特征的统计量,实现了线性复杂度的注意力计算。改进后的 TSSAIFI 模块如图 4 所示。

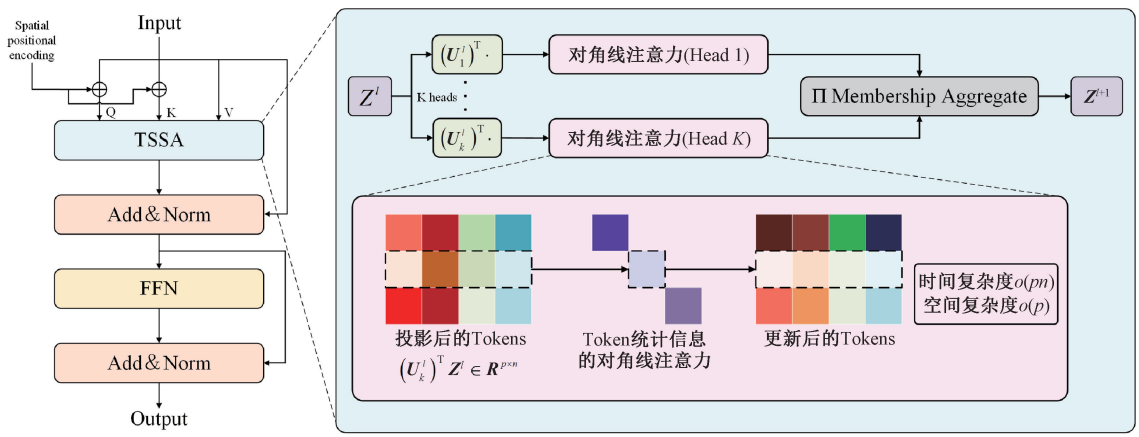


图 4　TSSAIFI 结构

Fig. 4　Structural diagram of TSSAIFI

在 TSSAIFI 中,将输入特征通过 3 个线性映射转换为 $\boldsymbol{Q}$(query)、$\boldsymbol{K}$(key)和 $\boldsymbol{V}$(value)3 个矩阵后进入 TSSA 模块。TSSA 模块将输入 Token 特征 $\boldsymbol{Z}^l$(l 为层数)通过低秩矩阵 $\boldsymbol{U}_k^l$ 投影到低维子空间,投影后的 Token

为$(\boldsymbol{U}_k^l)^{\mathrm{T}}\boldsymbol{Z}^l \in \boldsymbol{R}^{p\times n}$，$k$ 是注意力头的数量，p 为投影后的维度，n 为输入 Token 的数量，然后计算投影后特征的二阶矩统计量来建模数据的内在结构从而捕获令牌间的相互关系，将结果进行聚合，再经过层归一化处理，通过前馈网络(feed forward neural，FFN)对这些特征进行进一步的处理和增强，最后再层归一化得到最终输出。与传统 Transformer 不同的是 TSSA 不计算 Token 之间的成对相似度，而是通过统计特征来实现注意力机制，大大降低了计算复杂度。

原模型中 MHSA 的计算时间复杂度为 $O(pn^2)$，空间复杂度为 $O(n^2)$，而 TSSA 的计算时间复杂度为 $O(pn)$，空间复杂度为 $O(p)$，计算复杂度从二次方级别降低到线性级别，在处理大量高维 Token 时，计算效率得到显著提升。在 Long-Range Arena 基准测试中，TSSA 在序列长度 $N=4\ 096$ 时的推理速度比标准 Transformer 快 3.2 倍，内存占用减少 87%。实验表明，TSSA 相较于原模型的 MHSA 大幅减少模型参数量和冗余计算的同时性能上也能预期相媲美，模型效率显著提升。

2.3 RetBlockC3 模块

原模型中 RepC3 模块中的 RepConv 仅依赖局部卷积，无法有效捕捉长距离依赖，为增强空间信息捕获能力并降低计算复杂度，将其换成 RMT(retentive networks meet vision transformers)[19] 模型中的 RetBlock，RetBlockC3 模块结构如图 5 所示。RetNet(retentive network)引入基于距离的衰减机制用于序列建模，最初用于自然语言处理，提供显式时序先验。在计算机视觉中，此衰减机制扩展为基于曼哈顿距离的空间衰减，用于建模图像空间关系。

RetBlock 的设计结合了曼哈顿自注意力机制(MaSA)和其分解形式，通过曼哈顿距离计算空间衰减矩阵，动态调整注意力权重，引入显式空间先验，提供更丰富的空间位置信息，优于 RepConv 的固定卷积核。在二维空间中，目标 Token 与周围 Token 之间的注意力分数会根据曼哈顿距离衰减，越远的 Token 注意力分数衰减越大。这种机制使得模型能够更好地感知空间关系，同时保留全局信息。

曼哈顿自注意力机制公式如下：

$$\mathrm{MaSA}(\boldsymbol{X}) = (\mathrm{Softmax}(\boldsymbol{QK}^{\mathrm{T}}) \odot \boldsymbol{D}^{2d})\boldsymbol{V} \tag{1}$$

式(1)表明，MaSA 在计算注意力时，首先通过 Softmax 对原始注意力矩阵进行归一化，随后应用基于曼哈顿距离的空间衰减矩阵 $\boldsymbol{D}^{2d}$，矩阵 $\boldsymbol{D}$ 的每个元素被重新定义为基于 Token 对之间曼哈顿距离的指数衰减形式。

$$D_{nm}^{2d} = \gamma^{|x_n - x_m| + |y_n - y_m|} \tag{2}$$

式(2)表明，两个 Token 间的空间距离由横向坐标差 $|x_n - x_m|$ 与纵向坐标差 $|y_n - y_m|$ 之和(即曼哈顿距离)决定。通过指数衰减系数 γ，模型对不同距离的 Token 赋予不同的注意力权重，距离越近的 Token(如相邻像素块)衰减程度越小，对当前 Token 的影响权重越高；距离越远的 Token 衰减程度越大，权重逐渐降低。这种设计将显式的空间位置信息融入自注意力机制，既保留了全局依赖建模能力，又通过距离相关的衰减模式增强了局部特征的表达优先级。

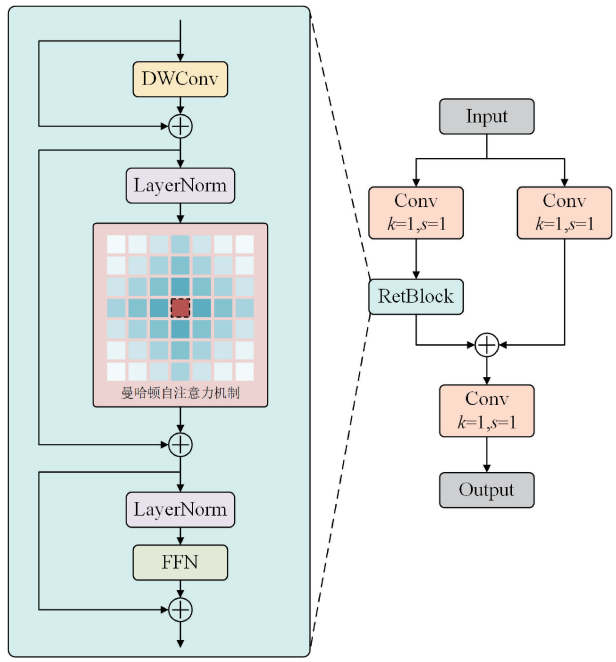


图 5 RetBlockC3 结构

Fig. 5 Structural diagram of RetBlockC3

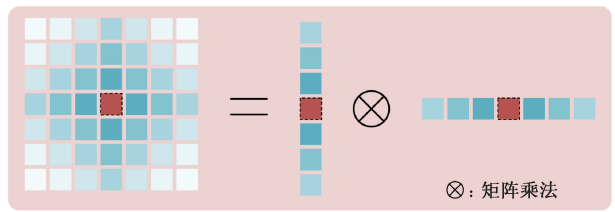


图 6 分解后的曼哈顿自注意力

Fig. 6 Decomposed Manhattan self-attention

为降低自注意力机制的二次计算复杂度，将自注意力与空间衰减矩阵沿图像的水平和垂直方向进行解耦，如图 6 所示。在保持曼哈顿距离空间先验的完整性前提下，通过分别计算行方向与列方向的注意力权重，将全局依赖建模的复杂度从二次方优化至线性，分解形式计算如式(3)~(5)所示，其中 $D_{nm}^H = \gamma^{|y_n - y_m|}$，$D_{nm}^W = \gamma^{|x_n - x_m|}$ 表示 Token 间的水平和垂直距离。

$$Attn_H = \mathrm{Softmax}(\boldsymbol{Q}_H\boldsymbol{K}_H^{\mathrm{T}}) \odot \boldsymbol{D}^H \tag{3}$$

$$Attn_W = \mathrm{Softmax}(\boldsymbol{Q}_W\boldsymbol{K}_W^{\mathrm{T}}) \odot \boldsymbol{D}^W \tag{4}$$

$$\mathrm{MaSA}(X) = Attn_H(Attn_W \boldsymbol{V})^{\mathrm{T}} \tag{5}$$

分解后的形式在保持与原始全矩阵形式相等的各向同性感受野的同时，显著降低了计算开销，尤其适用于高分辨率图像的密集预测任务。为进一步增强曼哈顿自注意力的局部表达能力，引入 DWConv，最后输出表示为：

$$X_{\mathrm{out}} = MaSA(\boldsymbol{X}) + LCE(\boldsymbol{V}) \tag{6}$$

2.4　边界框回归(GIoU)损失函数优化

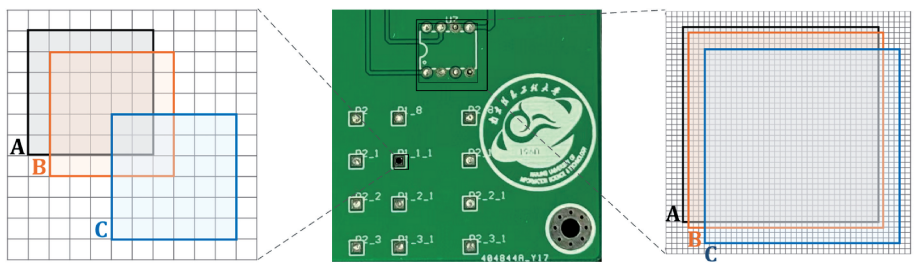
损失函数用于评估模型预测结果与实际结果之间的偏差，其值越低表示预测结果的准确性越高，即预测值与真实值越接近。RT-DETR 损失函数由分类损失和边界框回归损失组成，原模型中 GIoU Loss 通过衡量预测框与真实框的 IoU 和最小外接矩形面积优化定位精度，但忽略了边界框自身形状和尺度对回归结果的影响，从而限制了模型的优化和泛化能力以及可能引起算法过拟合。因此，本文使用了一种新的 FSN Loss 来提高目标检测的精度和定位的准确度，增强了其在边界框回归中的准确性和鲁棒性。

IoU 指标在目标检测任务中表现出显著的尺度敏感性差异。如图 7 所示，针对 6×6 pixels 的微小目标，亚像素级位置偏移即可引发 IoU 值的剧烈衰减(0.53→0.06)，而同等偏移量对 36×36 pixels 常规目标的 IoU 影响较小(0.90→0.65)。这种尺度相关的非线性特性导致传统 IoU 度量在跨尺度标签分配中存在双重局限性：首先，微小目标的 IoU 敏感性使其正负样本特征空间高度耦合，导致模型参数更新梯度混淆，显著降低网络收敛稳定性；其次，基于固定阈值的匹配机制在跨尺度场景下难以维持空间一致性，引发特征表征与几何定位的解耦偏差。因此引入一种基于 Wasserstein 距离的边界框相似性度量新指标，首先通过二维高斯分布建模构建边界框的概率表示，进而设计归一化 Wasserstein 距离(normalized Wasserstein distance，NWD)[20]来量化高斯分布间的相似性。该方法的主要优势在于突破了传统 IoU 对几何重叠区域的依赖性，在无重叠或极小重叠情境下仍能保持有效的相似性评估能力，其公式如式(7)、(8)所示。

$$D = \sqrt{(x_c - x_c^{gt})^2 + (y_c - y_c^{gt})^2 + \frac{(w - w^{gt})^2 + (h - h^{gt})^2}{weight^2}} \tag{7}$$

$$NWD = \mathrm{e}^{-\frac{D}{C}} \tag{8}$$

式中：D 是距离度量；(x_c, y_c) 和 (x_c^{gt}, y_c^{gt}) 分别是预测框和真实框的中心点坐标；w 和 h 是预测框的宽和高；w^{gt} 和 h^{gt} 是真实框的宽和高，如图 8 所示，$weight$ 和 C 是与数据集相关的常数，针对本文数据集特性选取 $weight = 2$，$C = 6$。



$IoU = \frac{|A \cap B|}{|A \cup B|} = 0.53 \Rightarrow IoU = \frac{|A \cap C|}{|A \cup C|} = 0.06$

(a) 小目标
(a) Small goals

$IoU = \frac{|A \cap B|}{|A \cup B|} = 0.90 \Rightarrow IoU = \frac{|A \cap C|}{|A \cup C|} = 0.65$

(b) 正常尺度目标
(b) Regular goals

图 7　NWD 示意图

Fig. 7　Schematic diagram of NWD

为更好地处理边界框回归中形状和尺度因素对回归结果的影响，利用 Shape-IoU[21]的思想构建 Shape-NWD 损失函数，如式(9)~(11)所示。

$$D_{\mathrm{Shape}} = \sqrt{w_w \times (x_c - x_c^{gt})^2 + h_h \times (y_c - y_c^{gt})^2 + \frac{(w - w^{gt})^2 + (h - h^{gt})^2}{weight^2}} \tag{9}$$

$$NWD_{\mathrm{Shape}} = \mathrm{e}^{-\frac{D_{\mathrm{Shape}}}{C}} \tag{10}$$

$$L_{\mathrm{Shape-NWD}} = 1 - NWD_{\mathrm{Shape}} \tag{11}$$

式中：w_w 与 h_h 分别为水平方向与垂直方向的权重系数，使得在水平和垂直方向上可以根据形状因素对这种偏移进行不同程度的加权。

考虑到困难样本和简单样本分布对边界框回归的影响，引进能够关注不同的回归样本的 Focaler-IoU[22]损失函数来提高检测器在不同检测任务中的性能，将其与 Shape-NWD 损失函数结合形成新的损失函数 FSN Loss。

Focaler-IoU 使用线性区间映射的方法重构 IoU 损

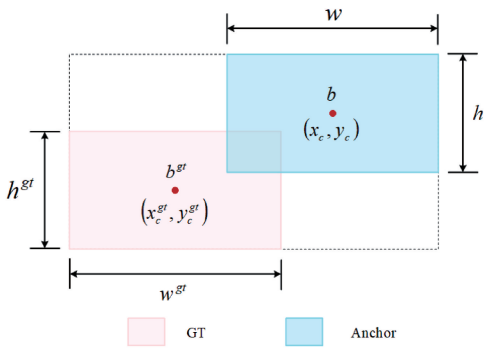


图 8 Shape-NWD 示意图

Fig. 8 Schematic diagram of Shape-NWD

失,其公式如式(12)、(13)所示。

$$IoU^{Focaler} = \begin{cases} 0, & IoU < d \\ \dfrac{IoU - d}{u - d}, & d \leqslant IoU \leqslant u \\ 1, & IoU > u \end{cases} \tag{12}$$

$$L_{Focaler-IoU} = 1 - IoU^{Focaler} \tag{13}$$

式中:IoU 是原始 IoU 值,$[d,u] \in [0,1]$,通过调整 d 和 u 的大小使 $IoU^{Focaler}$ 关注不同的回归样本。

最终边界框回归损失 FSN Loss 如式(14)所示。

$$L_{Focaler-Shape-NWD} = \alpha \times L_{Focaler-IoU} + \beta \times L_{Shape-NWD} = \alpha \times (1 - IoU^{Focaler}) + \beta \times (1 - e^{-\frac{D_{Shape}}{C}}) \tag{14}$$

式中:α 和 β 为权重参数,约束条件为 $\alpha + \beta = 1$。

3 实验结果与分析

3.1 实验环境搭建与参数设置

1)实验环境搭建

本文实验环境如表 1 所示。

表 1 实验环境配置

Table 1 Experimental environment configuration

名称	参数
CPU	Intel-i7-13700F
GPU	NVIDIA GeForce RTX 4080 16 G
RAM	32 G
操作系统	Windows 10
编程语言	Python 3.9.21
学习框架	Pytorch 1.13.1
GPU 加速库	CUDA 11.6

2)参数设置

为了确保实验变量的一致性,统一采 640×640 的输入尺寸,采用 AdamW 优化器进行参数的优化,模型训练迭代次数 epochs 为 150 次,权重衰减系数为 0.000 5,beta1 参数为 0.9,批量大小 batch size 设置为 16,初始学习率为 0.000 1,使用余弦退火(cosine annealing)策略动态调整学习率。为确保消融试验与对比试验具备更强的说服力,在后续的消融试验过程中,始终采用一致的超参数,而在针对不同模型展开的比较试验里,依据各模型的特性,运用其默认的超参数设置。

3.2 数据集准备

本文实验所使用的数据集为北京大学智能机器人开放实验室公开的 PCB 缺陷数据集,缺陷类型分为 6 类,漏孔(missing_hole)、鼠咬(mouse_bite)、开路(open_circuit)、短路(short)、毛刺(spur)以及余铜(spurious_copper)。由于该数据集只有 693 张缺陷图像,样本数量较少,容易产生过拟合,故本文通过随机翻转、灰度化、添加噪声、高斯模糊、锐化等数据增强方式将数据集扩增到 7 623 张图片,数量如图 9 所示,将处理后的数据集按 8∶1∶1 的比例随机划分训练集、验证集与测试集。

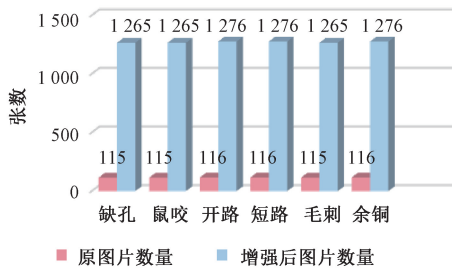


图 9 PCB 数据集数量

Fig. 9 Number of PCB datasets

3.3 评估指标

实验采用平均精度均值和精确率与召回率的调和平均数 F1 分值作为精度评价指标,使用帧率(FPS)计算模型的实时检测性能,这 3 个指标数值越大检测性能越好;而模型的轻量化效果由参数量(Params)和计算量(GFLOPs)这两个指标体现,数值越小越轻量。

精确率(precision,P)表示在检测出的全部目标中识别正确的比例,计算公式如式(15)所示。

$$P = \frac{TP}{TP + FP} \tag{15}$$

召回率(recall,R)表示正确预测为正的占全部实际为正的比例,计算公式如式(16)所示。

$$R = \frac{TP}{TP + FN} \tag{16}$$

式中:TP(true positive)为将实际为正的样本正确预测为正的数量;FP(false positive)为错误地将实际为负的样本分类为正的数量;FN(false negative)为未能检测到实际为正的样本。

平均精度值(average precision,AP)是 P-R 曲线的积

分，即曲线与坐标系所围面积，其公式为：

$$AP = \int_0^1 P(R)\,\mathrm{d}R \tag{17}$$

平均精度均值（mean average precision，mAP）是 n 个不同类别的 AP 和的平均值，反映了网络整体缺陷检测的准确性，公式如式（18）所示。

$$mAP = \frac{1}{n}\sum_{i=0}^{n} AP_i \tag{18}$$

$mAP@0.5$ 是 IoU 阈值为 0.5 时的 mAP，$mAP@0.5$：0.95 为不同 IoU 阈值（从 0.5 ~ 0.95，步长 0.05）上的 mAP。

F1 分值表示精准率与召回率的谐波平均值，如式（19）所示。

$$F1 = 2\frac{PR}{P+R} \tag{19}$$

3.4　消融实验

1）改进模块消融实验

为客观验证所提改进算法的有效性，以 RT-DETR-R18 为基准模型设计消融实验：将 G-StarNet、TSSAIFI、RetBlockC3、FSN 损失函数逐步整合至模型中（分别用 A、B、C、D 表示），同时将原算法作为对照组，评估不同模块对算法各项性能指标的影响。消融实验结果如表 2 所示。

表 2　消融实验结果对比

Table 2　Comparison of ablation experiment results

模型	mAP@0.5/%	F1 值/%	FPS/fps	Params/($\times10^6$)	计算量/($\times10^9$)
RT-DETR（Baseline）	95.1	94.2	115.3	19.9	57.0
RT-DETR+A	93.6	92.5	136.8	11.4	32.6
RT-DETR+A+B	94.5	93.9	137.1	11.1	31.9
RT-DETR+A+B+C	95.0	94.1	143.0	9.8	25.4
RT-DETR+A+B+C+D	96.7	95.3	142.6	9.8	25.4

由表 2 可知，将基准模型 RT-DETR 的主干网络替换为轻量化网络 G-StarNet 后，模型参数量和计算量显著降低，分别减少了 42.7% 和 42.8%，推理速度提高了 18.6%，但 mAP@0.5 也随之降低了 1.5%。为弥补精度的损失，在 AIFI 模块中引入通过统计特征来增强模型对关键信息关注度的 TSSA 自注意力机制，mAP@0.5 提升了 0.9%，参数量减少了 0.3×10^6，计算量减少了 0.7×10^9，FPS 略微变快。进一步集成 RetBlockC3 模块后，计算复杂度从二次方优化至线性，参数量减少了 1.3×10^6，计算量减少了 6.5×10^9，FPS 提高了 5.9 fps，mAP@0.5 提升了 0.5%。最后，把改进后模型中的损失函数替换为 FSN Loss，使其能更好地处理边界框回归中形状和尺度因素对回归结果的影响，关注不同的回归样本，mAP@0.5 提升了 1.7%，FPS 轻微下降，参数量和计算量几乎保持不变。综上所述，本文改进后的模型相比原模型平均精度均值提升了 1.6%，参数量减少了 50.8%，计算量减少了 55.4%，推理速度提高了 23.7%，这表明 STR-DETR 模型在实现轻量化的同时各项指标都具有正向效果。

2）边界框回归损失函数消融实验

为深入验证边界框回归损失函数改进策略的有效性，本文将原网络模型所采用的 GIoU 损失函数，与 NWD、Shape-NWD、Focaler-Shape-NWD 3 种改进后的损失函数进行对比分析。如图 10 所示，相较于其他 3 种损失函数，采用 FSN 损失函数训练数据集时，模型梯度下降速率最快，收敛后的损失值达到最低水平。这一结果充分表明，FSN 损失函数能够有效加快模型收敛速度，提升模型拟合度，进而对模型整体性能优化有一定的助益。

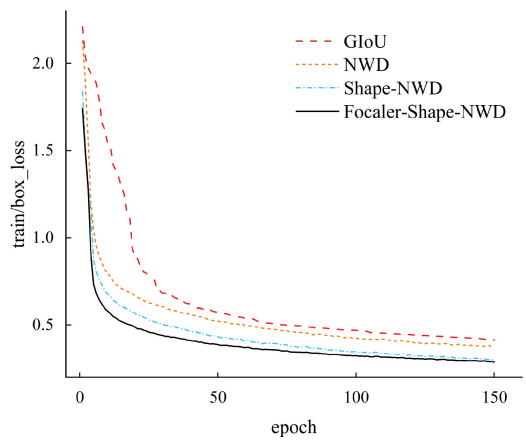


图 10　损失函数对比

Fig. 10　Comparison diagram of loss functions

3.5　对比实验

1）与其他算法对比实验

为进一步验证本文所提算法的优越性，将其与其他算法进行对比实验。鉴于两阶段算法检测实时性难以满足本研究实时检测需求，特选取实时性表现优异的 YOLO 系列目标检测算法作为对比对象。在统一实验环境与数据集条件下，选取 YOLOv5s、YOLOv5m、YOLOv8s、YOLOv8m、YOLOXs 等主流算法进行对比研究，具体实验结果如表 3 所示。

表 3 主流算法对比

Table 3 Comparison of mainstream algorithms

模型	mAP@0.5/%	FPS/fps	Params/(×10^6)	计算量/(×10^9)
YOLOv5s	89.3	156.0	7.2	16.5
YOLOv5m	94.6	97.4	21.2	48.3
YOLOv8s	92.5	128.7	11.2	28.8
YOLOv8m	96.9	76.4	25.9	78.9
YOLOXs	91.6	132.5	9.3	26.8
RT-DETR	95.1	115.3	19.9	57.0
STR-DETR	96.7	142.6	9.8	25.4

由表 3 可知,本文提出的 STR-DETR 模型的检测精度 mAP@0.5 达到 96.7%,仅比最高的 YOLOv8m 模型低 0.2%;在实时性方面,STR-DETR 的 FPS 为 142.6 fps,虽略低于 YOLOv5s,但显著高于其他高精度模型,且其推理速度是 YOLOv8m 的 1.87 倍,验证了其高效性;模型复杂度方面,STR-DETR 的参数量仅为 9.8×10^6,较 RT-DETR 压缩了 50.8%,同时计算量低至 25.4×10^9,较 YOLOv8s 和 YOLOXs 分别减少 11.8%和 5.2%。值得注意的是,尽管 YOLOv5s 的计算量更小,但其检测精度与 STR-DETR 差距显著,而 STR-DETR 在仅比体量最小的 YOLOv5s 增加 2.6×10^6 参数量的情况下实现了精度与速度的双重提升。综上所述,本文模型在实现轻量化的同时保持了检测精度,适用于实时性与高精度并重的工业检测场景。

2)轻量化主干网络对比

在严格控制其他参数保持一致的条件下,以 RT-DETR-R18 作为基准网络,将 FasterNet、MobileNet V4、VanillaNet、GhostNet V2、Shufflenet V2 等主流轻量化网络分别替换基准网络的主干来进行对比实验,实验结果如表 4 所示。

表 4 不同轻量化主干网络对比

Table 4 Comparison of different lightweight backbone networks

Backbone	mAP@0.5/%	Params/(×10^6)	计算量/(×10^9)
FasterNet	93.3	13.6	35.9
MobileNet V4	93.5	13.9	40.7
VanillaNet	92.8	12.5	34.2
GhostNet V2	93.7	14.6	39.5
Shufflenet V2	93.0	12.7	34.9
G-StarNet	93.6	11.4	32.6

由表 4 可知,G-StarNet 与 FasterNet、MobileNet V4、VanillaNet、GhostNet V2、Shufflenet V2 等主流轻量化网络在平均精度方面检测效果表现相近,然而在模型复杂度量化分析中,G-StarNet 展现出一定的优势,其参数量与计算量均低于其余对比模型。这表明在保持特征提取能力相当的条件下,G-StarNet 通过更高效的网络架构设计,实现了模型轻量化目标。基于上述实验结果,本文选择 G-StarNet 作为改进模型的主干网络,旨在平衡检测精度与计算资源消耗,为后续算法优化提供高效的基础框架。

3.6 边缘设备测试

为进一步验证本文改进的轻量化模型部署在边缘设备上的可行性,选取 NVIDIA 开发的嵌入式计算模块 Jetson TX2 作为模型部署的边缘计算设备,搭建平台如图 11 所示,进行网络模型相关环境配置及依赖项安装,将 PyTorch 模型转换为 onnx 格式,通过 TensorRT 对其优化,生成高效的推理引擎,提高模型在 Jetson TX2 上的推理速度,改进前后模型的实验结果如表 5 所示。

图 11 Jetson TX2 平台

Fig. 11 Jetson TX2 platform

表 5 边缘设备部署结果

Table 5 Edge device deployment results

模型	mAP@0.5/%	FPS/fps	Params/(×10^6)	Weights/MB
RT-DETR	95.1	9	19.9	40.5
STR-DETR	96.6	17	9.8	18.1

由表 5 可知,改进前后模型在边缘计算设备上呈现的精度与 PC 端几乎保持一致,而原模型 FPS 仅为 9 fps,改进后提升了 88.9%,能够在单位时间内处理更多图像,满足实时检测的需求,参数量与模型权重相较于原模型均减少了 1/2 以上,部署时比原模型需要更少的计算资源和存储空间,适用于资源受限环境。综上,本文所提轻量化模型可在计算资源受限的边缘设备上有效部署,且凭借其较快的推理速度,契合工业生产实时缺陷检测需求。

3.7 可视化分析

为实现 STR-DETR 模型性能的直观呈现,本文从测试集中选取 6 种不同类型缺陷的图像,采用 GradCAM++方法对改进前后的模型检测效果进行热力图可视化分析。GradCAM++是 GradCAM 的改进版本,通过引入高阶梯度加权机制,更精准地定位深度学习模型在图像中的关键区域,提升可视化结果的可解释性和细粒度特征捕捉能力。实验结果如图 12 所示,通过对比可以发现,本文提出的模型在整体置信度上超过了原模型,并有效减

少了毛刺的误检现象,充分证实了本文方法的显著效果。

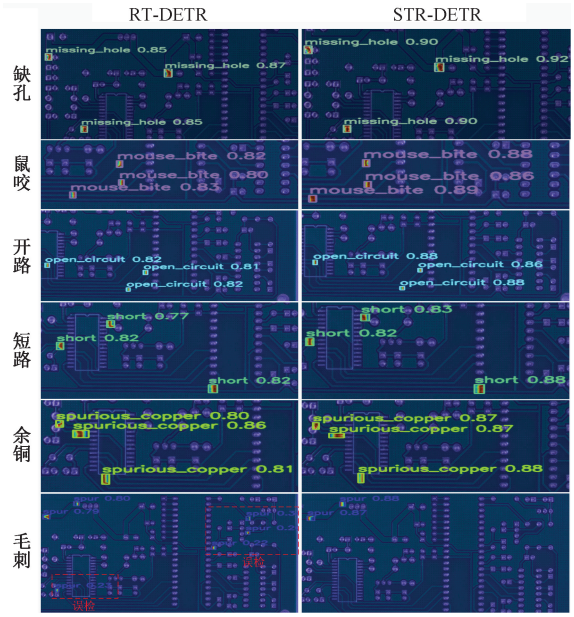


图 12　STR-DETR 与原模型检测效果对比

Fig. 12　Comparison of the detection effects between STR-DETR and the original model

4　结　论

本文针对 PCB 缺陷检测中高计算复杂度和参数冗余导致的边缘设备部署瓶颈,提出一种基于 STR-DETR 的轻量化检测方法。通过将 GConv 引入轻量级网络 StarNet,构建了 G-StarNet 来替换原有主干网络,在保留模型高效性的同时显著降低了参数量与计算复杂度。在 AIFI 模块中引入基于统计学特征的 TSSA 自注意力机制代替多头自注意力,由复杂的成对相似度计算转变为高效的使用二阶矩统计量,降低了计算开销。利用 RetBlock 重新设计 RepC3 模块,通过距离相关的衰减模式增强了局部特征的表达优先级,计算复杂度从二次方优化至线性。选取结合 3 种损失函数优势构成新的损失函数 FSN Loss,通过归一化 Wasserstein 距离来量化高斯分布间的相似性,并改善了形状和尺度因素、困难与简单样本分布对边界框回归结果的影响。该算法通过优化网络结构与特征处理流程,在保持检测精度的同时显著降低了模型参数量和计算量,这一设计旨在解决边缘计算场景下的资源约束问题,为实时目标检测技术在终端设备上的高效部署提供了新的解决方案。未来将深入研究更复杂的缺陷类型、小样本学习问题,以及提升在复杂环境下 PCB 表面缺陷检测的性能。

参考文献

[1] LING Q, ISA N A M. Printed circuit board defect detection methods based on image processing, machine learning and deep learning: A survey[J]. IEEE Access, 2023, 11: 15921-15944.

[2] ZHENG J, SUN X, ZHOU H, et al. Printed circuit boards defect detection method based on improved fully convolutional networks[J]. IEEE Access, 2022, 10: 109908-109918.

[3] GIRSHICK R, DONAHUE J, DARRELL T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014: 580-587.

[4] GIRSHICK R. Fast R-CNN[C]. Proceedings of the IEEE International Conference on Computer Vision, 2015:1440-1448.

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

[6] 陈仁祥,詹赞,胡小林,等. 基于多注意力 Faster RCNN 的噪声干扰下印刷电路板缺陷检测[J]. 仪器仪表学报,2021,42(12):167-174.
CHEN R X, ZHAN Z, HU X L, et al. Defect detection of printed circuit boards under noise interference based on multi-attention faster RCNN[J]. Chinese Journal of Scientific Instrument, 2021, 42(12): 167-174.

[7] REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: Unified, real-time object detection[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2016: 779-788.

[8] YASEEN M. What is YOLOv8: An in-depth exploration of the internal features of the next-generation object detector[J]. ArXiv preprint arXiv:2408. 15857, 2024.

[9] WANG A, CHEN H, LIU L, et al. YOLOv10: Real-time end-to-end object detection[J]. ArXiv preprint arXiv:2405. 14458, 2024.

[10] TIAN Y, YE Q, DOERMANN D. YOLOv12: Attention-centric real-time object detectors[J]. ArXiv preprint arXiv:2502. 12524, 2025.

[11] WEI L, DRAGOMIR A, DUMITRU E, et al. SSD: Single shot multibox detector[C]. Proceedings of the European Conference on Computer Vision. Springer, 2016: 21-37.

[12] 张莹,邓华宣,王耀南,等. 基于多通道特征融合学习的印制电路板小目标缺陷检测[J]. 仪器仪表学报,2024,45(5):10-19.
ZHANG Y, DENG H X, WANG Y N, et al. Small target defect detection of printed circuit boards based on multi-

channel feature fusion learning [J]. Chinese Journal of Scientific Instrument, 2024, 45(5): 10-19.

[13] 王军,伍毅,陈正超. 基于 SMT-YOLOv8 的 PCB 缺陷检测研究[J]. 电子测量技术,2024,47(11):131-137.
WANG J, WU Y, CHEN ZH CH. Research on PCB defect detection based on SMT-YOLOv8 [J]. Electronic Measurement Technology, 2024, 47(11): 131-137.

[14] CARION N, MASSA F, SYNNAEVE G, et al. End-to-end object detection with transformers [C]. Proceedings of the European Conference on Computer Vision. Springer, 2020: 213-229.

[15] 吴斌斌,张礼华,刘军伟,等. 基于改进的 EP-RTDETR 小目标 PCB 表面缺陷检测[J]. 制造技术与机床, 2025(3):139-148.
WU B B, ZHANG L H, LIU J W, et al. Small-target pcb surface defect detection based on improved EP-RTDETR [J]. Manufacturing Technology & Machine Tool, 2025 (3): 139-148.

[16] ZHAO Y, LYU W, XU S, et al. Detrs beat YOLOs on real-time object detection[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024: 16965-16974.

[17] MA X, DAI X, BAI Y, et al. Rewrite the stars[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024: 5694-5703.

[18] WU Z, DING T, LU Y, et al. Token statistics transformer: linear-time attention via variational rate reduction[J]. ArXiv preprint arXiv:2412.17810, 2024.

[19] FAN Q, HUANG H, CHEN M, et al. RMT: Retentive networks meet vision transformers[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024: 5641-5651.

[20] WANG J, XU C, YANG W, et al. A normalized Gaussian Wasserstein distance for tiny object detection[J]. ArXiv preprint arXiv:2110.13389, 2021.

[21] ZHANG H, ZHANG S. Shape-IOU: More accurate metric considering bounding box shape and scale[J]. ArXiv preprint arXiv:2312.17663, 2023.

[22] ZHANG H, ZHANG S. Focaler-IOU: More focused intersection over union loss[J]. ArXiv preprint arXiv: 2401.10525, 2024.

作者简介

陈枫赟,2023 年于南京信息工程大学获得学士学位,现为南京信息工程大学硕士研究生,主要研究方向为机器视觉、目标检测。
E-mail: 772771670@qq.com

Chen Fengyun received her B. Sc. degree from Nanjing University of Information Science and Technology in 2023. Now she is a M. Sc. candidate at Nanjing University of Information Science and Technology. Her main research interests include machine vision and object detection.

李鹏(通信作者),2008 年于西安交通大学获得博士学位,现为南京信息工程大学教授,博士生导师,中国仪器仪表学会气象仪器分会理事,主要研究方向为深度学习、图像处理。
E-mail: lipengnuist@163.com

Li Peng(Corresponding author) received his Ph. D. degree from Xi'an Jiaotong University in 2008. Now he is a professor and Ph. D. supervisor at Nanjing University of Information Science and Technology, and a CIS-MIS director. His main research interests include deep learning and image processing.