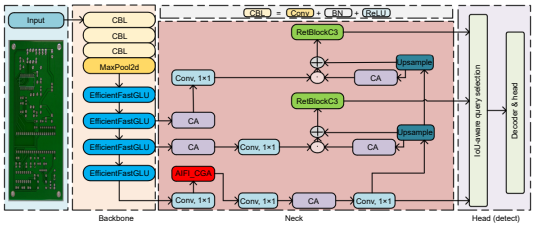
OEE | 光电工程 Opto-Electronic Engineering

Article

2026 年 2 月, 第 53 卷, 第 2 期

基于 LightRT-DETR 算法的印刷电路板缺陷检测



希治远[1,3], 梅　婷[1,3], 林珊玲[2,3*], 林志贤[1,2,3], 林坚普[2,3], 吕珊红[2,3], 郭太良[1,3]

[1]福州大学物理与信息工程学院, 福建 福州 350116;
[2]福州大学先进制造学院, 福建 泉州 362000;
[3]中国福建光电信息科学与技术创新实验室, 福建 福州 350116

摘要: 针对工业生产中印刷电路板 (printed circuit board, PCB) 缺陷检测精度和实时性不高的问题, 本文提出了一种基于 LightRT-DETR 的 PCB 缺陷检测算法。首先, 设计了一种高效的主干网络 EfficientFastGLU-18, 该网络通过引入部分卷积 (partial convolution, PConv) 和卷积门控线性单元 (convolutional gated linear unit, convolutional GLU) 机制来降低 ResNet-18 架构的计算复杂度和参数量, 同时保持较好的特征提取能力。其次, 采用级联分组注意力机制替代了 RT-DETR 中的多头注意力机制, 通过特征分割与级联, 增强输入特征的多样性, 并降低不必要的计算开销。最后, 提出了一种高效选择性特征金字塔网络 (efficient selective feature pyramid network, ES-FPN), 其主要由特征选择模块和串联 RetBlockC3 的特征融合模块构成, 增强对细微缺陷的检测能力的同时减少计算冗余, 从而实现更高效的实时缺陷检测。实验结果表明, 本文改进的模型的 mAP50 达到了 98.3%, mAP50-95 值较基准模型 RT-DETR 提高了 11.1%, GFLOPs 下降了 34.7%, Params 降低了 35.3%, FPS 提升了约 5.1%。这些优化使得 LightRT-DETR 模型在移动设备或嵌入式系统中更加适用。

关键词: 缺陷检测; 印刷电路板; 主干网络; 特征金字塔网络; 计算冗余

The English title and abstract appear at the end of the paper.

DOI: 10.12086/oee.2026.250117 | CSTR: 32245.14.oee.2026.250117 | 中图分类号: TP391 | 文献标识码: A

希治远, 梅婷, 林珊玲等. 基于 LightRT-DETR 算法的印刷电路板缺陷检测 [J]. 光电工程, 2026, **53**(2): 250117

Xi Z Y, Mei T, Lin S L et al. Printed circuit board defect detection based on LightRT-DETR algorithm[J]. *Opto-Electron Eng*, 2026, **53**(2): 250117

1 引　言

印刷电路板 (PCB) 制造行业在电子工业的版图中占据着举足轻重的地位。然而, 在 PCB 生产过程中, 缺陷检测环节仍然面临着巨大挑战, 传统的人工检测效率低下、准确性不足, 难以满足现代电子制造业对高质量和高效率的严格要求。此外, 像短路、开路、鼠咬等缺陷的检测会引起视觉疲劳, 容易导致误分类[1]。为了克服这些问题, 深度学习算法在 PCB 缺陷检测领域展现出巨大潜力。

目前, PCB 缺陷检测的深度学习算法主要分为两大类: 基于传统深度卷积神经网络 (convolutional neural networks, CNN) 的目标检测算法和基于 Transformer 架构的目标检测算法。基于 CNN 的缺陷检测算法又可细分为两类, 单阶段算法与两阶段算法。单阶段算法主要包含 YOLO[2] 系列、RetinaNet[3] 和单点多盒检测 SSD[4]

收稿日期: 2025-04-19; 修回日期: 2025-06-23; 录用日期: 2025-06-23; 网络出版日期: 2026-02-13
基金项目: 国家重点研发计划 (2021YFB3600603)
*通信作者: 林珊玲, sllin@fzu.edu.cn
版权所有©2026 中国科学院光电技术研究所

系列算法。两阶段算法主要包括 R-CNN[5]、Fast R-CNN[6]、Faster R-CNN[7] 和 Mask R-CNN[8]。近年来，研究者们对这些算法进行了多方面的改进，Lian 等[9] 利用 mask R-CNN 进行缺陷检测，并在全连接 CNN 中添加几何注意力引导的 mask 分支，以提高模型的效率。Li 等[10] 将 Faster R-CNN 和 YOLOv2 融合，提高了 PCB 双列直插式组件的检测效率并降低了误检测率。Xin 等[11] 针对 PCB 的缺陷检测，提出了一种增强的 YOLOv4 模型。该方法引入马赛克数据增强策略和 Mish 激活函数，提高了小目标检测能力。Xie 等[12] 使用多阶段残差混合注意力模块 MRHAM 改进特征学习，并通过 K-means++聚类优化锚框，显著提高了检测准确性。Wu 等[13] 提出 GSC YOLOv5 模型，通过 Ghost Conv 和 Ghost Bottleneck 实现轻量化，并引入双重注意力机制提高检测精度和效率。Wang 等[14] 提出基于轻量级深度学习的缺陷检测网络，使用 CSPDarkNet 作为 YOLOX 主干网络，并融入协调注意力机制，提高了小型 PCB 缺陷检测能力。陶志勇等[15] 提出了 CViT-Net 网络，通过融合局部与全局特征提升缺陷检测性能，并设计轻量与标准两种结构以兼顾精度与效率。

基于 Transformer 的目标检测算法在处理全局信息、捕捉细节特征和建立序列关系方面展现出独特优势，相较于传统的 CNN 方法更具竞争力。基于此，Carion 等[16] 提出了一种完全端到端的检测器 DETR，采用集合预测方法，巧妙地规避了非极大值抑制 (non-maximum suppression, NMS) 后处理步骤的需求。该模型运用匈牙利算法来实现对象分类和回归，展现出优秀的性能。Zhao 等[17] 提出了 RT-DETR 模型，RT-DETR 使用一个高效的混合编码器取代了常规 Transformer 编码器。RT-DETR 分离处理不同尺度特征的内部交互和跨尺度融合，这一策略显著提升了模型的推理速度。此外，RT-DETR 采用了基于不确定性最小化的查询选择机制，这种方法通过精确优化不确定性，为解码器提供高质量的初始查询，从而提高了整体检测精度。

但是，现阶段 RT-DETR 也有些许不足之处，针对这些不足本文将对其进行优化。1) ResNet[18] 作为 RT-DETR 的主干网络结构需要大量的计算资源和存储资源，尤其是在处理高分辨率图像时，这使得它在资源受限的环境中应用受到限制。所以本文设计出一种新的主干网络 EfficientFastGLU-18 优化其带来的计算负担。2) RT-DETR 中的基于注意力的内部尺度特征交互 AIFI 采用了多头注意力机制 MHA[19]，需要计算多个注意力头，并且多个注意力头可能会捕捉到相似的信息，导致参数冗余，本文将采用级联分组注意力 CGA[20]，通过级联和分组的方式，提高了模型的计算效率。并且对于漏孔和鼠咬等需要精细局部特征的缺陷，CGA 的逐层特征提取能力可以更准确地捕捉细微的形状变化和边缘特征。3) RT-DETR 中的路径聚合网络 (path aggregation network, PANet) 只能完全融合相邻层级的特征。这限制了模型对多样化缺陷的识别能力。同时，在融合高低层级特征时，PANet 无法专注于特定的语义信息。这对检测如杂刺和杂散等细微缺陷尤为重要，因为这些缺陷可能表现为微小的异常纹理或颜色变化。此外，双向金字塔结构中存在大量路径和间接信息交互，显著增加了模型的计算量。基于此本文提出了 ES-FPN 网络解决上述问题。结合上述改进策略，本文设计了 LightRT-DETR 模型，这是一种适合执行 PCB 缺陷检测任务的轻量级实时目标检测模型。

2 LightRT-DETR 模型介绍

RT-DETR 是一种基于 Transformer 架构的实时端到端目标检测器，本文选择了 RT-DETR-R18 作为基础模型，该模型在检测精度和模型复杂度之间取得了良好平衡。如图 1 所示，RT-DETR-R18 的网络结构主要包括 ResNet18 主干网络用于特征提取，高效混合编码器 (包含 AIFI 模块和 CCFM) 负责多尺度特征转换，以及解码器部分。

本文从特征提取、Transformer 结构、多尺度特征转换等方向出发，对 RT-DETR-R18 模型进行模块与结构改进，提出 LightRT-DETR 新方法，如图 2 所示。LightRT-DET 模型由主干网络、混合编码器和带辅助预测头的变压器解码器组成。利用轻量级的主干网络输出特征作为编码器的输入。混合编码器通过基于 CGA 的尺度内交互机制与一种高效选择性特征金字塔网络将多尺度特征转换为一系列图像特征。随后使用 IoU 感知从编码器输出序列中选择固定数量的图像特征作为解码器的初始对象查询。最后，使用带有辅助预测头的解码器迭代优化对象查询以生成方框和置信度分数。

3 相关工作

3.1 骨干网络 EfficientFastGLU-18

在现代计算机视觉任务中，神经网络的计算效率和

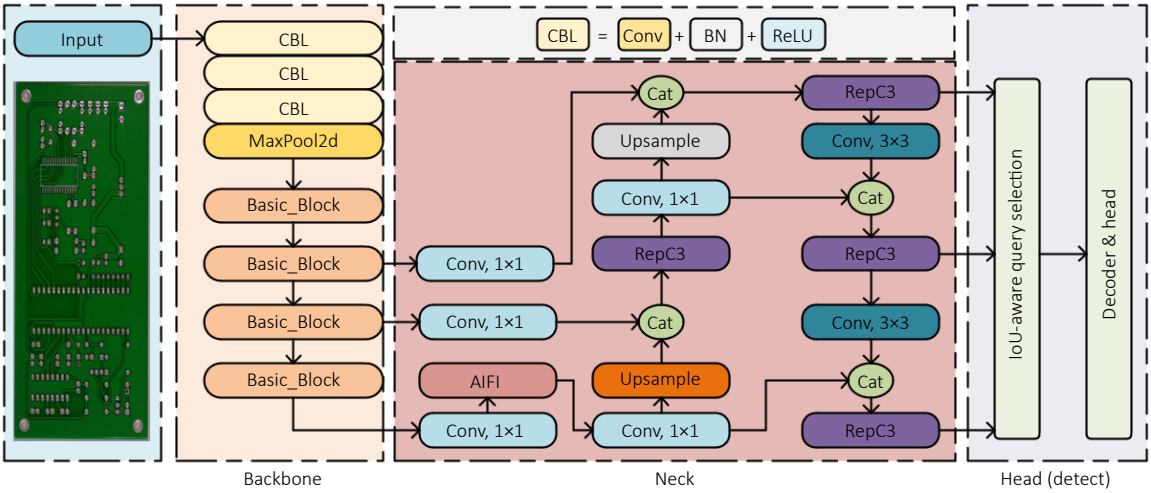


图 1 RT-DETR-R18 模型网络结构

Fig. 1 Network architecture of the RT-DETR-R18 model

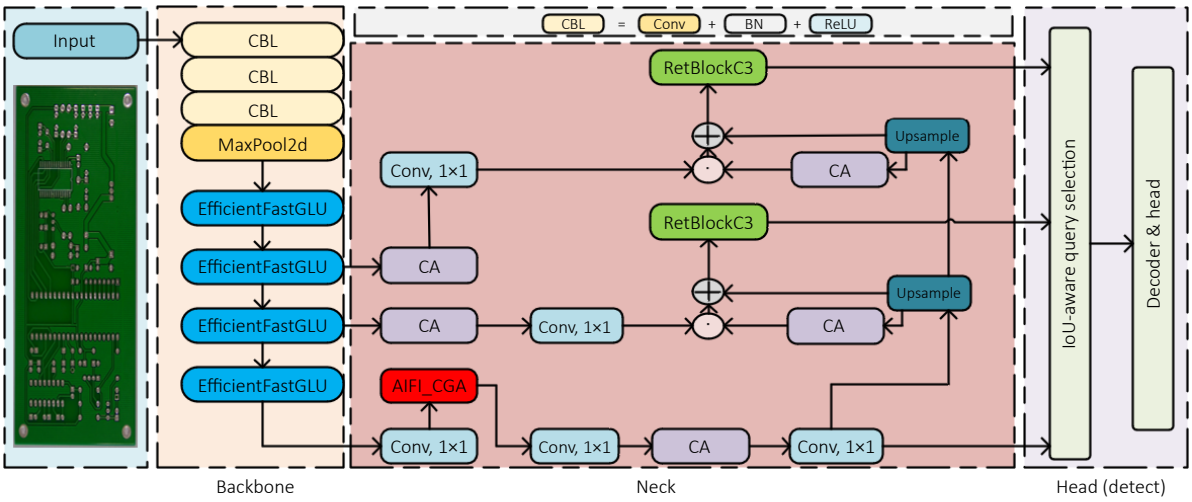


图 2 LightRT-DETR 模型网络结构

Fig. 2 Network architecture of the LightRT-DETR model

参数数量是至关重要的性能指标。针对 RT-DETR 模型在处理高分辨率图像时资源消耗高的问题，本文首先对 ResNet-18 中的 BasicBlock 进行优化，引入 FasterBlock 来替代传统的卷积块。随后，为了进一步提高特征提取能力，本文引入了 Convolutional GLU，对 FasterBlock 进行二次改进。这种改进后的模块被命名为 EfficientFastGLU，如图 3 所示，其最终实现了在保持较高精度情况下提升计算效率和减少参数量的目标。

EfficientFastGLU 中的核心之一是部分卷积。如图 4 所示，可以同时减少计算量和内存访问，更高效地提取空间特征。PConv 利用特征图的冗余性，选择性地仅对部分输入通道 (c_p) 进行卷积操作，而其余通道则使用恒等映射。假设输入和输出特征图的通道数量相同，这种方法不会丧失一般性。与传统的卷积操作对所有通道 (c) 进行卷积相比，PConv 能够显著减少浮点运算次数，有效降低计算复杂度和内存访问量，如式 (1) 所示。

$$r_{\mathrm{F}}=\frac{h\times w\times k^2}{h\times w\times k^2\times c^2}\times\frac{c^2}{r^2}, \tag{1}$$

式中：r_{F}为 PConv 的浮点运算次数与普通卷积浮点运算次数的比值；h、w 分别为输入特征图的高和宽；k 为卷积核大小；r 为通道缩减因子，其取值为 0.25，对于该取值安排，在第 4 节中进行了详细展示说明。由此可见，PConv 的计算量仅为普通卷积的 1/16，内存访问数量仅仅为普通卷积的 1/4。在主干特征提取网络中引入

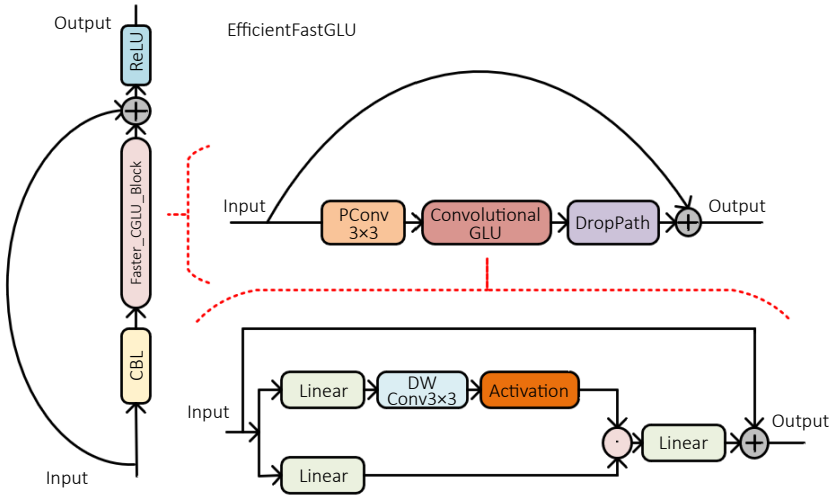


图 3 EfficientFastGLU 架构

Fig. 3 EfficientFastGLU architecture

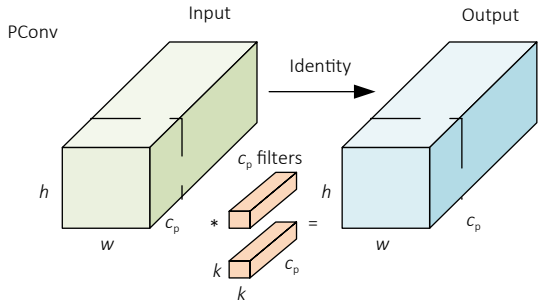


图 4 Pconv 网络结构图

Fig. 4 Network structure diagram of Pconv

PConv 算子能够显著降低计算量和内存访问量，从而使骨干网络轻量化并提高模型的推理速度。

此外，EfficientFastGLU 具有强大的特征提取能力是因为其引入了 Convolutional GLU。这是对传统压缩-激励 (squeeze-and-excitation, SE) [21] 机制的改进。如图 5 所示，Convolutional GLU 作为一种创新的通道混合器，克服了 SE 机制中全局平均池化导致的局限性。它由两个线性变换组成，一个通过非线性激活函数作为门控信号，另一个作为值分支。这使得每个 token 都能基于其邻近特征获得定制的门控信号，实现更精细和灵活的通道注意力调整。

EfficientFastGLU-18 网络通过结合 FasterBlock 和 Convolutional GLU，在保持与卷积前馈网络 (convolutional feed-forward network, ConvFFN)相似参数量的前提下，显著降低了计算复杂度。这种设计不仅确保了模型的高效性和训练稳定性，还在 ResNet-18 的基础上实现了性能的显著提升。

3.2 基于 CGA 的尺度交互机制

本文使用了级联分组注意力机制 CGA 来代替 AIFI 中的多头注意力机制 (multi-head attention, MHA)，以此来提高模型性能和降低冗余参数。这种注意力机制可以表示为

$$\tilde{X}_{ij} = Attn\left(X_{ij}W_{ij}^{\boldsymbol{Q}},\ X_{ij}W_{ij}^{\boldsymbol{K}},\ X_{ij}W_{ij}^{\boldsymbol{V}}\right), \tag{2}$$

$$\tilde{X}_{i+1} = Concat\left[\tilde{X}_{ij}\right]_{j=1:h} W_i^{\boldsymbol{P}}, \tag{3}$$

式中：$\boldsymbol{Q}$ 是查询向量，代表当前位置想要查询的信息。$\boldsymbol{K}$ 是键向量，代表所有位置的信息索引，用于和 $\boldsymbol{Q}$ 计算相似度，得到注意力权重。$\boldsymbol{V}$ 是值向量，代表所有位置的实际内容，注意力权重会作用在 $\boldsymbol{V}$ 上，得到最终的加权输出。$\boldsymbol{P}$ 是投影权重矩阵，用于对多头注意力拼接后的特征做线性变换，完成特征融合与维度映射。i 代表网络层级索引，j 代表多头注意力中的注意力头索引。第 j 个头计算 X_{ij} 的自注意力，X_{ij} 是输入特征 X_i 的第 j 个分割部分，即 $X_i=[X_{i1}, X_{i2}, \cdots, X_{ih}]$，且 $1\leqslant j\leqslant h$。h 是头的总数，$W_{ij}^{\boldsymbol{Q}}$、$W_{ij}^{\boldsymbol{K}}$、$W_{ij}^{\boldsymbol{V}}$ 是将输入特征分割映射到不同子空间的投影层，而 $W_i^{\boldsymbol{P}}$ 是一个线性层，将连接后的输出特征投影回与输入一致的维度。

CGA 机制通过避免传统方法中对完整特征的全面依赖，转而采用特征分割与级联分组策略。具体来说，CGA 机制首先将输入特征分成多组，每组包含若干个通道。然后，将不同的注意力头通过级联连接起来，形

成级联注意力结构，如图 6 所示。在该结构中，每个注意力头只关注一组输入特征，并将注意力计算的输出传递给下一个注意力头。最后，级联结构的输出特征通过跨层连接与其他层的特征相连，增强特征表达。

$$X'_{ij} = X_{ij} + \tilde{X}_{i(j-1)},\quad 1 < j < h, \tag{4}$$

式中：X'_{ij}是第 j 个输入分割X_{ij}和通过式 (2) 计算的第 (j-1) 个头输出$\tilde{X}_{i(j-1)}$的相加结果。在计算自注意力时，它取代X_{ij}作为第 j 个头的新输入特征。此外，在 q 投影之后应用了另一个 token 交互层，使自注意力能够共同捕捉局部和全局关系，进一步增强特征表示。CGA 机制通过使用分组、级联、跨层连接等方式，可以增强输入特征的多样性，减少计算冗余，提高模型性能和计算效率。

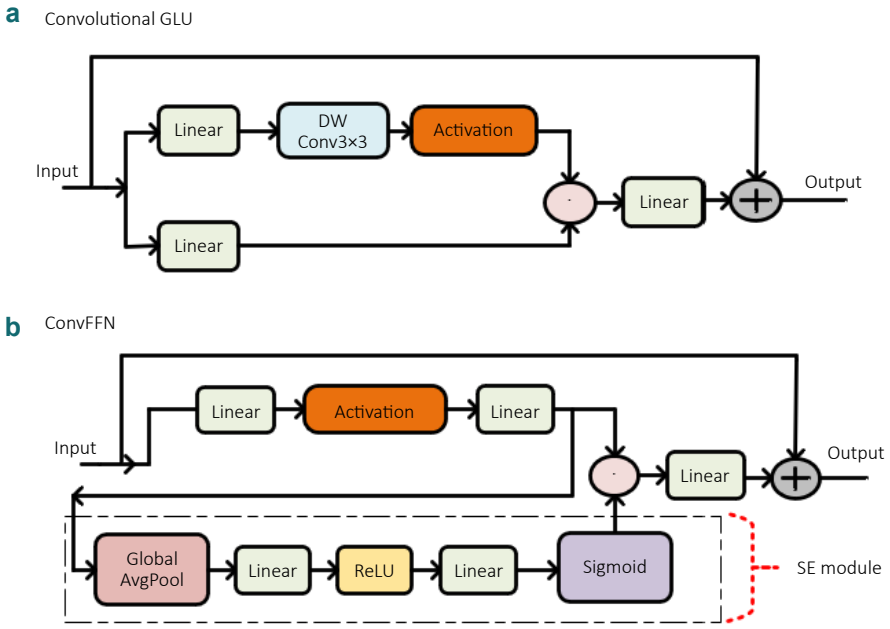


图 5 Convolutional GLU 模块与 ConvFFN 模块结构对比图。(a) Convolutional GLU 模块；(b) ConvFFN 模块

Fig. 5 Structural comparison between Convolutional GLU module and ConvFFN module. (a) Convolutional GLU module; (b) ConvFFN module

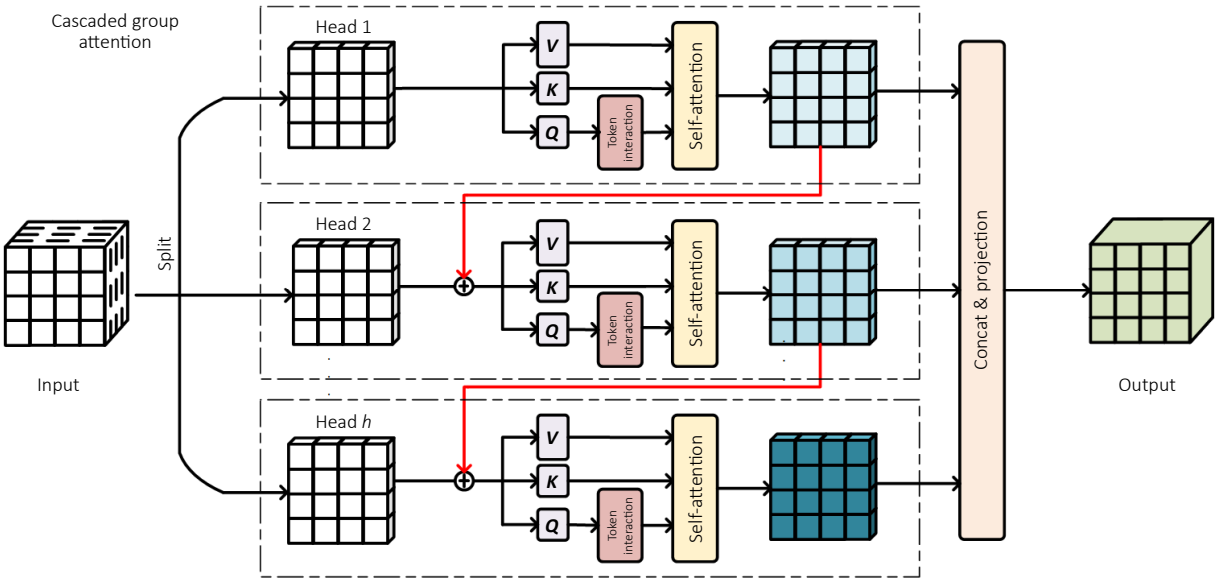


图 6 CGA 注意力机制

Fig. 6 CGA attention mechanism

3.3 ES-FPN 特征金字塔网络

RT-DETR 模型中，CCFM 使用了 PANet 来实现多尺度特征的融合。虽然这种方法通过自顶向下和自底向上的双向金字塔结构来增强特征的交互和定位能力，但存在以下三个主要问题：1) PANet 只能完全融合相邻层级的特征。对于不相邻的层级，特征信息需要通过递归的方式间接获取。例如，漏孔和鼠咬可能需要更细粒度的特征，而开路和短路可能需要更大范围的上下文信息。仅依赖相邻层级的特征融合可能会限制模型对这些多样化缺陷的识别能力。2) 在融合高低层级特征时，PANet 无法专注于特定的语义信息。这对检测如杂铜和杂散等细微缺陷尤为重要，因为这些缺陷可能表现为微小的异常纹理或颜色变化，需要模型能够捕捉非常细致的特征。3) 双向金字塔结构中存在大量路径和间接信息交互，显著增加了模型的计算量。在 PCB 生产线上进行实时缺陷检测时，检测速度直接影响生产效率，因此需要一个更高效的特征融合机制。为解决上述问题，本文提出一种高效选择性特征金字塔网络，其网络结构图如图 7 所示。其中，S3、S4、S5 分别表示主干网络第 3、4、5 阶段输出的特征图，N3、N4、N5 分别表示经过初步特征选择后的多尺度特征图。ES-FPN 架构主要是由特征选择模块与串联 RetBlockC3 的特征融合模块构成。

3.3.1 特征选择模块

通道注意力机制 (channel attention, CA) 模块和维度匹配 (dimension matching, DM) 模块在特征选择过程中协同发挥着关键作用。具体而言，CA 模块，如图 8 所示，首先对输入特征图 $f_{\text{in}} \in \mathbf{R}^{C\times H\times W}$ 进行处理，其中 C 表示通道数，H 和 W 分别表示特征图的高度和宽度。该特征图分别经过全局平均池化和全局最大池化操作，获得两组全局描述信息，随后将两组池化特征进行融合，并通过一个 Sigmoid 激活函数生成每个通道的权重 $f_{\text{CA}} \in \mathbf{R}^{C\times 1\times 1}$。这些权重与原始特征图逐通道相乘，实现对重要通道的增强和无效通道的抑制，从而有效提取每个通道中的关键信息。接下来，经过 CA 模块加权的特征图输入到 DM 模块，DM 模块通过 1×1 卷积、上采样或下采样等操作对特征的空间分辨率和通道数进行调整，使不同尺度的特征在空间和通道维度上实现一致性匹配。通过这种先通道加权、后维度对齐的融合逻辑，特征选择模块能够充分挖掘多尺度特征的有效信息，并为后续的特征融合和检测任务提供高质量的特征基础。

3.3.2 串联 RetBlockC3 的特征融合模块

特征融合模块通过选择性特征融合机制，使用高级特征作为权重来过滤低尺度特征中的重要信息。高级特征经过转置卷积和双线性插值操作进行尺度调整来匹配低尺度特征的尺寸，然后利用高级特征作为注意力权重，筛选出低尺度特征中有用的信息。并且通过将最高层特征 N5 不经处理直接输入解码器，最大程度地保留了其内部蕴含的丰富语义信息。为解决特征融合过程中的计算负担问题，本文从文献 [22] 中汲取灵感，串联了 RetBlockC3 模块，如图 9 所示，其中 $C1/2$ 表示输入特征通道数 $C1$ 的一半，$C2$ 表示模块输出的特征通道数。这一策略确保了筛选后的重要语义信息能够有效融合，同时提高计算效率。RetBlockC3 主要是利用曼哈顿自注意力机制 MaSA，减少了计算复杂度，直接降低了处理大规模数据集时的内存消耗。

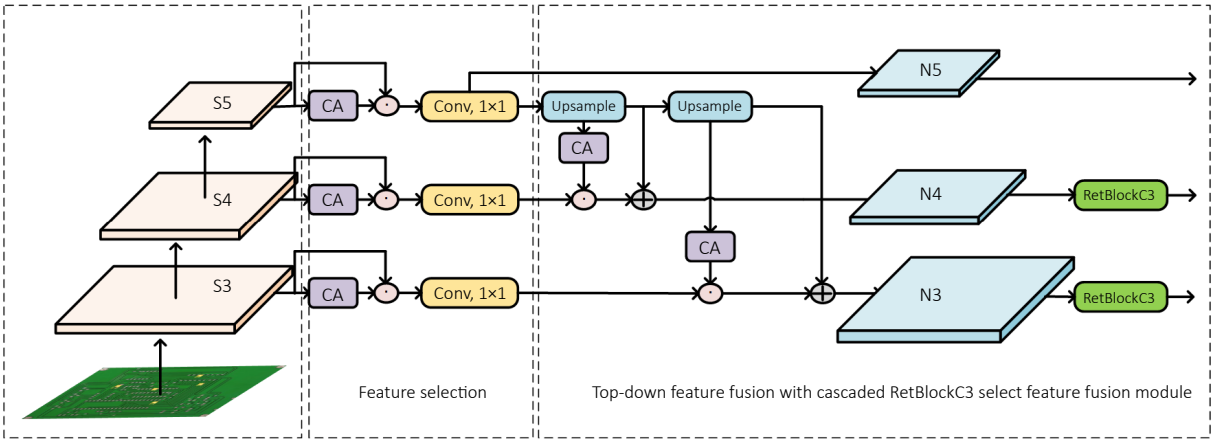


图 7 ES-FPN 架构

Fig. 7 ES-FPN architecture

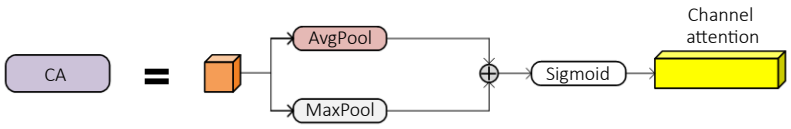


图 8 CA 模块架构

Fig. 8 CA Module architecture

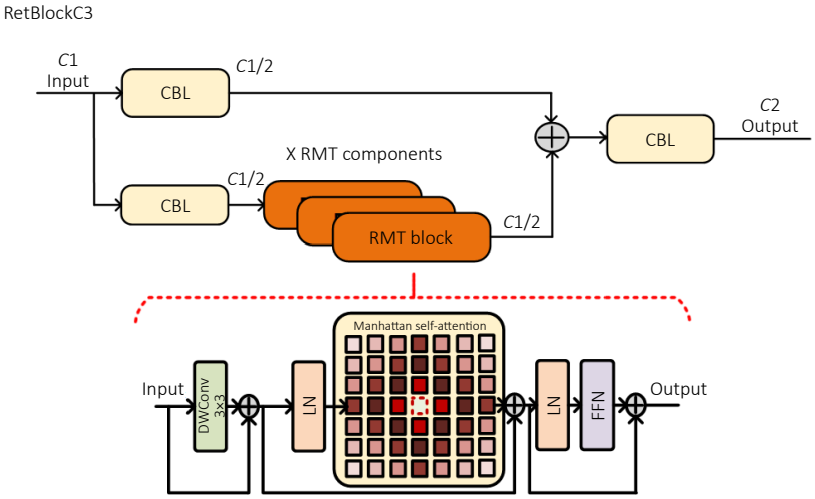


图 9 RetBlockC3 模块架构

Fig. 9 RetBlockC3 module architecture

MaSA 将垂直方向上的注意力分数 (经过值矩阵 $\boldsymbol{V}$ 加权后) 进行转置，并与水平方向上的注意力分数相乘，得到最终的注意力输出 $MaSA(\boldsymbol{X})$。这种分解和重组的方式，既保留了自注意力机制的全局感知能力，又通过分解降低了计算成本。

$$\gamma_i = \log\left(1 - 2^{-\left(2+\frac{4\cdot i}{num_heads}\right)}\right), \tag{5}$$

$$Attn_{\mathrm{H}} = Softmax(\boldsymbol{Q}_{\mathrm{H}}\boldsymbol{K}_{\mathrm{H}}^{\mathrm{T}}) \odot \boldsymbol{D}^{\mathrm{H}}, \tag{6}$$

$$Attn_{\mathrm{W}} = Softmax(\boldsymbol{Q}_{\mathrm{W}}\boldsymbol{K}_{\mathrm{W}}^{\mathrm{T}}) \odot \boldsymbol{D}^{\mathrm{W}}, \tag{7}$$

$$MaSA(\boldsymbol{X}) = Attn_{\mathrm{H}}(Attn_{\mathrm{W}}\boldsymbol{V})^{\mathrm{T}}, \tag{8}$$

式中：$\boldsymbol{D}^{\mathrm{H}}_{nm} = \gamma_i^{|y_n - y_m|}$，$\boldsymbol{D}^{\mathrm{W}}_{nm} = \gamma_i^{|x_n - x_m|}$，其中 n、m 为像素点的索引；γ 为衰减因子；i 为注意力头的索引，i=0, 1, ⋯, num_heads−1，由于 RT-DETR-R18 共有 8 个注意力头，所以 num_heads 为 8；$\boldsymbol{Q}_{\mathrm{H}}$、$\boldsymbol{K}_{\mathrm{H}}$分别为水平方向的查询和键；$\boldsymbol{Q}_{\mathrm{W}}$、$\boldsymbol{K}_{\mathrm{W}}$分别为垂直方向的查询和键；$\boldsymbol{V}$ 为值矩阵；$\boldsymbol{D}^{\mathrm{H}}$、$\boldsymbol{D}^{\mathrm{W}}$分别为水平方向和垂直方向的位置衰减矩阵。

4 实验与结果分析

4.1 数据集介绍

本文使用了北京大学智能机器人开放实验室发布的开源数据集，数据集中共有 693 张图片涵盖了 6 种 PCB 缺陷类型，包括漏孔、鼠咬、开路、短路、杂铜和杂散，每张图片的分辨率为 2777 pixel×2138 pixel。每张图片包含 3~6 个缺陷。训练集、验证集和测试集划分比为 7:1:2。具体数据集信息如表 1 所示。

表 1 数据集信息

Table 1 Dataset information

Defect type	Number of images	Number of defects
Missing hole	115	449
Mouse bite	115	411
Open circuit	116	453
Short circuit	116	433
Spurious copper	116	473
Spur	115	443

4.2 实验环境与参数设置

实验平台的硬件环境为 GPU 为 RTX 3090，显存为 24 GB，CPU 为 Intel Xeon Platinum 8362，内存为 45 GB，操作系统及软件环境为 Ubuntu 18.04、Cuda 11.1、Python 3.8、PyTorch 1.9.0。送入模型训练的图片大小均被设置为 640×640，batch_size 设置为 4，epoch 设置为 200，采用自适应权重衰减优化器 AdamW，初始学习率

为 0.0001，lrf 设置为 1.0，优化器权重衰减系数和动量系数分别设置为 0.0001 和 0.9，颜色空间增强方面，hsv_h 设置为 0.015，hsv_s 设置为 0.7，hsv_v 设置为 0.4，几何变换增强方面，translate设置为 0.1，fliplr 设置为 0.5，scale 设置为 0.5。

4.3 评价指标

为了分析模型对 PCB 板的缺陷检测效果的检测效果，选择常用的精确率 (precision, *P*)、召回率 (recall, *R*)、平均精度均值 (mAP) 作为衡量算法性能的主要指标。选择模型参数量 (Params)、计算量 (GFLOPs)、检测速度 (FPS) 等参数作为衡量模型推理性能的指标。

$$R=\frac{TP}{TP+FN}, \tag{9}$$

$$P=\frac{TP}{TP+FP}, \tag{10}$$

$$AP=\int_0^1 P(R)\mathrm{d}R, \tag{11}$$

$$mAP=\frac{\sum_{j=1}^{C}AP_j}{C}. \tag{12}$$

其中：*TP* 是真正例，即模型正确识别的正样本数量。*FN* 是假反例，即模型错误判定为负样本的正样本数量。*FP* 是假正例，即模型错误判定为正样本的负样本数量。

4.4 结果分析

4.4.1 同其他主干网络改进的对比

为了验证主干网络改进的有效性，本文在 RT-DETR 的基础上，选择了不同的轻量化方案作为对比。对比实验在验证集上的实验结果如表 2 所示。

由表 2 可知，EfficientFastGLU-18 在目标检测任务中具有较好性能。与其他方案相比，EfficientFastGLU-18 不仅降低了计算复杂度，GFLOPs 从 57.0 降至 46.2，参数量也从 19.8 M 减少到 15.4 M，还保持了较高的推理速度，达到 57.2 f/s，最关键的是该模型实现了最高的检测精度，其 mAP50 达到 96.9%。这种全面的性能提升源于其创新的网络设计：部分卷积显著提高了计算效率，Convolutional GLU 增强了特征提取能力，而整体网络结构的优化则在效率和精度之间取得了良好平衡。相比之下，其他模型如 EfficientViT 和 RepViT 虽然在轻量化方面表现出色，但检测精度分别降至 90.0% 和 88.4%，可能是过度追求效率而牺牲了特征提取能力。Fasternet 和 MobileNet_v4 虽然在 FPS 方面领先，分别达到 64.9 和 64.0 f/s，但 mAP50 相对较低，分别为 94.4% 和 91.5%，表明它们可能在复杂特征提取方面存在不足。EfficientFormer_v2 虽然保持了较高的 mAP50，但 FPS 明显降低，反映了在特征提取效率和精度平衡上的挑战。这些对比凸显了 EfficientFastGLU-18 在目标检测任务中实现高效率、高精度平衡的良好表现。

4.4.2 引入部分卷积合理性分析

本文在 EfficientFastGLU-18 主干网络中引入了部分卷积，部分卷积 Pconv 中的一个重要参数就是通道缩减因子 *r*。*r* 过小会导致信息损失，影响特征表达能力，*r* 过大会引入冗余计算，降低模型效率，对此本文对其进行了系统验证。结果如表 3 所示，通道缩减因子 *r*=0.250 是精度与效率的最佳平衡点。实验数据表明，相比 *r*=0.125，*r*=0.250 使 mAP50 提高了 1.2%，计算量仅微增 1.5%；相比 *r*=0.500，*r*=0.250 保持相同的检测精度，同时计算量减少 5.7%，推理速度提高 7.1%。这一选择不仅在理论上平衡了信息保留与计算效率，也在实践中证明其能够有效捕获关键特征信息，同时避免冗余计算。

由于在 EfficientFastGLU-18 中引入了 PConv，所以可能会导致特征图表达能力减弱，并且引入梯度非连续

表 2 主干网络的改进策略对比

Table 2 Comparison of improvement strategies for backbone networks

Base model	Backbone network improvement	mAP50/%	GFLOPs	FPS	Params/M
RT-DETR	Resnet18 (baseline)	96.2	57.0	56.4	19.8
RT-DETR	EfficientFormer_v2	95.7	29.5	44.7	11.8
RT-DETR	EfficientViT	90.0	27.2	32.1	10.7
RT-DETR	RepViT	88.4	36.4	45.5	4.4
RT-DETR	Fasternet	94.4	28.5	64.9	10.8
RT-DETR	MobileNet_v4	91.5	39.5	64.0	11.3
RT-DETR	**EfficientFastGLU-18**	**96.9**	**46.2**	**57.2**	**15.4**

注：加粗数据为本文改进策略。

表 3 EfficientFastGLU-18 采用不同的通道缩减因子 *r* 结果对比

Table 3 The results comparison of EfficientFastGLU-18 with different channel reduction factors *r*

Reduction ratio *r*	mAP50/%	GFLOPs	FPS
0.125	95.7	45.5	58.8
0.250	**96.9**	**46.2**	**57.2**
0.500	96.9	49.0	53.4

注：加粗数据为本文使用策略。

性或退化学习问题，由此本文对该问题进行了验证。验证结果如图 10 与图 11 所示，EfficientFastGLU-18 结构在特征表达能力、梯度连续性和学习稳定性方面均展现出显著优势。首先，在特征表达能力方面，该结构不仅未降低模型表现，反而显著提升了特征谱能量。图 11(b) 显示所有层的相似度都在 0.4~0.5 之间，说明 EfficientFastGLU-18 模型在这些层的特征分布与原始模型有一定差异，但仍有较高的相关性；图 11(c) 显示 EfficientFastGLU-18 模型在各层的特征谱能量均高于原始模型，尤其在第 4 层提升约 2 倍。同时，如图 10(b)、图 11(a) 显示，EfficientFastGLU-18 模型在 mAP50 性能指标上持续优于原始模型。其次，关于梯度非连续性问题，如图 10(a) 显示，EfficientFastGLU-18 模型的损失曲线平滑下降，无异常波动，训练过程无震荡现象，EfficientFastGLU-18 模型收敛更快。若存在梯度非连续性应表现为收敛不稳定或迟缓，但实际并未观察到该现象。同时，在退化学习问题方面，图 10(b) 显示 PConv 模型在训练初期即可达到较高性能，第 30 个 epoch 时

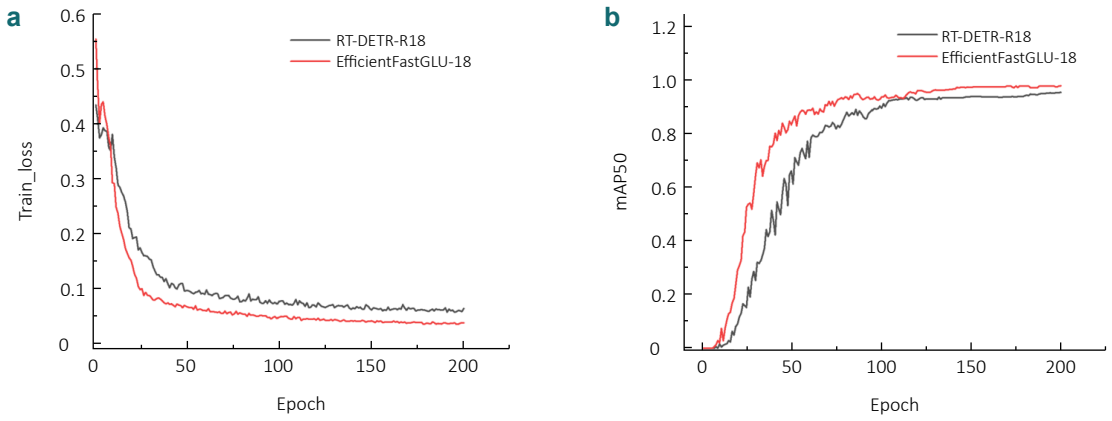


图 10 引入 PConv 前后的损失值图像与 mAP 图像对比。(a) 损失值变化的对比图；(b) mAP50 变化的对比图

Fig. 10 Comparison of loss values and mAP images before and after incorporating PConv. (a) Comparison image of loss value changes; (b) Comparison image of mAP50 changes

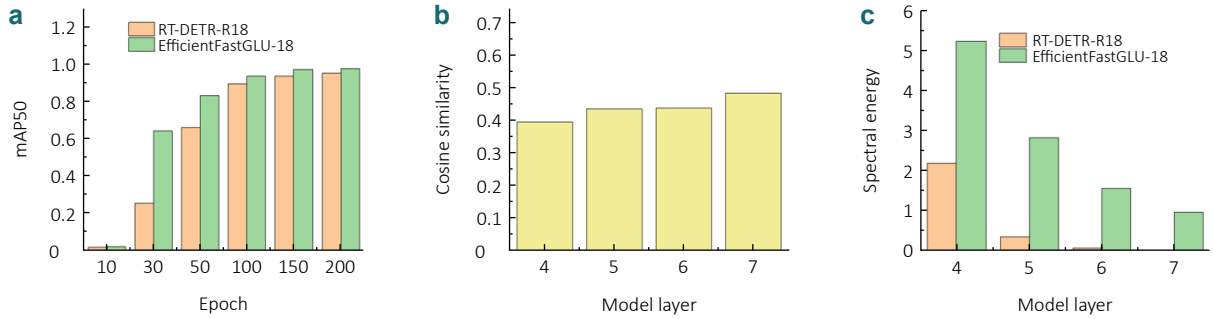


图 11 引入 PConv 前后的性能与特征分析。(a) mAP50 变化的对比柱状图；(b) RT-DETR-R18 与 EfficientFastGLU-18 的特征空间相似性；(c) RT-DETR-R18 与 EfficientFastGLU-18 的特征谱能量对比

Fig. 11 Performance and feature analysis before and after introducing PConv. (a) Comparison bar chart of mAP50 changes; (b) Feature space similarity between RT-DETR-R18 and EfficientFastGLU-18; (c) Feature spectral energy comparison between RT-DETR-R18 and EfficientFastGLU-18

mAP 提升高达 38.9%，并且在整个训练周期内各项指标持续领先，无性能退化迹象。EfficientFastGLU-18 的结构创新在于通道分离和门控机制的结合，一方面通过部分卷积高效提取局部特征，另一方面保留原始信息并增强特征选择能力，从而在 GFLOPs 从 57.0 降至 46.2、参数量从 19.8 M 减少到 15.4 M 的同时，保持了较高的推理速度 (57.2 f/s) 和最高 (96.9%) 的平均精度均值。综上，EfficientFastGLU-18 是一种理论合理、实验有效的高效结构优化方法，在提升计算效率的同时，显著增强了模型性能和泛化能力。

4.4.3 同其他注意力机制的对比

结合表 4 数据可以看出，CGA Attention 在目标检测任务中展现出显著的综合性能优势。与其他主流注意力机制相比，CGA 实现了最高的检测精度，其 mAP50 达到 97.7%。这一性能不仅明显优于基准模型 Baseline 的 96.2%，也超越了 Deformable Attention 的 95.2%，以及多种轻量化注意力机制。在计算复杂度方面，CGA 的 GFLOPs 为 57.0，与 Baseline 和 Deformable Attention 等方法基本一致，说明该机制并未引入额外的计算开销。模型参数量方面，CGA 的参数规模为 19.7 M，亦与其他方法保持在相同数量级。从推理速度来看，CGA 达到了 57.8 f/s，明显优于 Baseline 的 56.4 f/s 和 Deformable Attention 的 55.0 f/s，同时接近 Performer 的 58.7 f/s 和 Linformer 的 60.3 f/s。虽然部分轻量化机制在推理速度上稍具优势，但其精度普遍低于 CGA，特别是 Focused Linear Attention 和 HiLo Attention，在降低计算成本的同时，其检测精度也出现了明显下滑。综合而言，CGA Attention 通过引入创新的上下文引导机制，在不增加显著计算负担的前提下，有效提升了特征表达能力与目标检测精度。

4.4.4 同其他特征颈部结构优化对比

结合表 5 的数据可以看出，ES-FPN 在目标检测任务中展现出优异的综合性能。在检测精度方面，ES-FPN 实现了 97.4% 的 mAP50，显著优于基准特征融合

表 4 注意力机制的改进策略对比

Table 4 Comparison of attention mechanism improvements

Base model	Attention	mAP50/%	GFLOPs	FPS	Params/M
RT-DETR	MHA attention (baseline)	96.2	57.0	56.4	19.8
RT-DETR	Focused linear attention	94.9	56.7	56.7	19.2
RT-DETR	Performer attention	95.7	56.7	58.7	19.3
RT-DETR	Linformer attention	95.6	55.7	60.3	19.3
RT-DETR	HiLo attention	94.8	57.1	53.4	19.8
RT-DETR	Deformable attention transformer	95.2	57.2	55.0	19.8
RT-DETR	**CGA attention**	**97.7**	**57.0**	**57.8**	**19.7**

注：加粗数据为本文改进策略。

表 5 特征颈部的改进策略对比

Table 5 Comparison of feature neck improvement strategies

Base model	Neck network improvement	mAP50/%	GFLOPs	FPS	Params/M
RT-DETR	PANet (baseline)	96.2	57.0	56.4	19.8
RT-DETR	BIFPN	95.5	64.3	41.4	20.3
RT-DETR	HSFPN	96.4	53.3	50.5	18.1
RT-DETR	SDFM	95.8	61.9	44.4	21.3
RT-DETR	PSFM	95.2	71.2	37.7	22.4
RT-DETR	**ES-FPN**	**97.4**	**47.9**	**57.0**	**17.4**

注：加粗数据为本文改进策略。

结构 PANet 的 96.2%，同时也超过了其他主流方法，如 BIFPN 的 95.5%、HSFPN 的 96.4% 以及 SDFM 的 95.8%，充分体现了其在特征表达能力方面的显著优势。在计算效率方面，ES-FPN 的计算复杂度为 47.9 GFLOPs，在所有对比方法中最低，明显优于 BIFPN 的 64.3 和 PSFM 的 71.2。这表明 ES-FPN 能够在保持高检测精度的同时，有效降低计算开销。此外，其参数量仅为 17.4 M，为所有结构中最小的，进一步彰显该模型在轻量化设计方面的优势。ES-FPN 通过引入通道注意力机制，有效增强非相邻层级间的信息交互能力，提升对多样化缺陷的识别效果。同时，该结构能够突出关键语义通道，增强了对细微缺陷的检测能力。配合轻量化的 RetBlockC3 模块和 MaSA 注意力机制，ES-FPN 有效降低计算负担，提升推理速度，兼顾检测精度与效率。

4.4.5 消融实验

为了验证所提方法中各模块的有效性，本文进行了一系列消融实验，结果如表 6 所示。基准模型 M1 不包含任何提出的模块，而 M2 至 M7 分别添加了不同组合的模块。实验结果表明，各模块均对检测性能产生了积极贡献。在 M2 中单独引入 EfficientFastGLU-18 模块后，mAP50 提升了 0.7%，同时计算量减少 18.9%，参数规模下降 22.2%。这一结果充分证明该模块能够在提升模型精度的同时实现显著的轻量化优化。M3 所引入的 CGA Encoder 模块带来了 1.5% 的 mAP50 提升，表明增强模型对全局上下文信息的建模能力对于提高检测效果至关重要。M4 中加入的 ES-FPN 模块在提升检测精度 1.2% 的同时，计算量减少了 16.0%，体现出该结构在多尺度特征融合方面的高效性与实用性。在模块组合方面，M5 联合使用 EfficientFastGLU-18 与 ES-FPN，在检测精度和计算效率之间实现了良好平衡。M7 则结合了 CGA Encoder 与 ES-FPN，mAP50 达到 97.8%，为所有非完整模型中精度最高的组合，验证了其在特征增强与融合方面的互补优势。最终的完整模型 M8 整合了所有三个模块，实现了最佳的 98.3% 的 mAP50，同时保持了低计算复杂度仅为 37.2 的 GFLOPs 和 12.8 M 的小模型尺寸，推理速度达到 59.3 f/s。三大模块各有侧重，联合使用时能实现性能的全面提升，展现出明显的协同增益效应。

尽管前述消融实验系统性地评估了各改进模块对整体检测性能的提升效果，但在实际缺陷检测任务中，不同类型目标 (如小目标与大目标) 由于其特征分布和检测难度的差异，可能对改进策略的响应存在显著不同。为进一步验证各模块在不同目标尺寸下的适用性及优势，本文对各模块在小目标与大目标检测任务中的表现进行了细致分析。具体而言，依据小目标目前的通用定义将测试集中缺陷面积小于 32 pixel×32 pixel 的目标定义为小目标，大于该阈值的定义为大目标，并分别对两类目标进行检测性能评估，相关结果见表 7。

从结果可以看出，所提出的三大改进模块在提升大、小目标检测性能方面均展现出积极效果，但各自优化机制侧重点不同。EfficientFastGLU-18 模块所构建的模型 M2 通过引入轻量化深层注意力机制，有效提升了特征提取效率，尤其在小目标检测方面表现突出，其小目标 mAP50-95 由基础模型的 47.5% 提升至 51.8%。与此同时，大目标检测精度也得到一定提升，显示出对多尺度特征的良好适应能力。CGA Encoder 模块对应的模型 M3 进一步强化了局部空间建模与上下文理解能力，使大目标 mAP50 达到 98.4%，小目标的 mAP50-95 提升至 52.3%。该模块在增强全局感知和特征聚合方面的能力，显著提升了复杂目标结构的检测效果。ES-FPN 模块所构建的模型 M4 通过跨层次特征融合，有效整合了

表 6 消融实验结果

Table 6 Ablation experiment results

Model	EfficientFastGLU-18	CGA encoder	ES-FPN	mAP50/%	mAP50-95/%	GFLOPs	FPS	Params/M
M1	×	×	×	96.2	47.6	57.0	56.4	19.8
M2	√	×	×	96.9	51.8	46.2	57.2	15.4
M3	×	√	×	97.7	52.4	57.0	57.8	19.7
M4	×	×	√	97.4	52.4	47.9	57.0	17.4
M5	√	×	√	97.6	**53.0**	**37.2**	58.7	12.9
M6	√	√	×	96.9	52.0	46.3	58.0	15.2
M7	×	√	√	97.8	52.8	48.0	58.4	17.2
M8	√	√	√	**98.3**	52.9	**37.2**	**59.3**	**12.8**

注：×为不添加模块；√为添加模块； 加粗数据为最优值。

浅层高分辨率信息与深层语义表达，提升了大目标检测精度至 98.0%，小目标检测性能也达到 97.2%，其 mAP50-95 提高至 52.2%，表明其在多尺度融合方面具备良好效果。在此基础上，多个模块的组合进一步强化了模型性能。模型 M7 联合 CGA Encoder 与 ES-FPN，在大目标的 mAP50-95 上达到最高值 56.0%。最终模型 M8 整合了 EfficientFastGLU-18、CGA Encoder 与 ES-FPN 三大模块，实现了在大、小目标检测精度上的全面最优表现，大目标 mAP50 达到 98.5%，小目标 mAP50 达到 98.2%，小目标 mAP50-95 提升至 52.4%。

上述实验结果表明，三大模块具有良好的互补性。EfficientFastGLU-18 提升了不同尺度目标的快速识别能力，CGA Encoder 加强了小目标边缘细节与大目标结构理解，ES-FPN 有效平衡了小目标的精细定位与大目标的语义表达。最终集成模型在不同目标尺度下均实现最优或近最优性能，充分验证了所提方法的鲁棒性与广泛适应性。

4.4.6 同其他模型的对比

为了验证本文所提算法的优越性，本文选择了几种主流的目标检测算法进行对比，包括 YOLOv5l、YOLOv8m、YOLOv8l、YOLOv9m 和 YOLOv9e，同时也包括文献 [23] 提到的 EP-RTDETR、文献 [24] 提到的基于 Faster R-CNN 的改进模型和文献[25] 提到的 LIDD-YOLO，同时还涵盖了 WiseYOLOv7[26]、FPFM[27]、LDD-Net[28]、ExtendedFPN[29] 和 TD-Net[30] 等近年来代表性方法。本文所提 LightRT-DETR 算法在多个关键指标上均表现出色。实验结果如表 8 所示。首先，在检测精度方面，EP-RTDETR 缺乏对 PCB 缺陷检测任务的针对性设计，导致其 mAP50 比本文算法的提到的低 1.6%，并且在模型规模上处于明显劣势。尽管 WiseYOLOv7 的 mAP50-95 高于本文算法，但是其计算量是本文算法的 3.6 倍，具有较高计算复杂度。本文算法的优异性能主要源于三个关键创新：首先，引入 EfficientFastGLU-18 主干网络，有效降低了高分辨率 PCB 图像处理的资源需求；其次，采用级联分组注意力替代多头注意力机制，显著提高了参数效率，特别适合捕捉细微缺陷特征；最后，提出 ES-FPN 网络优化特征融合，增强了对多样化 PCB 缺陷的识别能力。总体来看，虽然 YOLOv 系列、WiseYOLOv7 等模型在通用目标检测任务中表现优异，但在 PCB 缺陷检测这一特定任务上，存在模型规模庞大、易过拟合或参数效率低等局限性。部分轻量级模型如 LIDD-YOLO、LDD-Net、TD-Net 等虽然模型小巧，但检测精度略显不足。

4.4.7 损失值与 mAP 对比

图 12 展示了 RT-DETR 与 LightRT-DETR 两种模型在训练过程中的性能对比，从图 12(a) 可以看出，两个模型的损失值均随训练轮数增加而迅速下降，且 LightRT-DETR 的损失下降速度更快，收敛速度优于 RT-DETR。在约 100 个 epoch 后，两个模型的损失值趋于稳定，但 LightRT-DETR 的损失始终低于 RT-DETR。图 12(b) 进一步显示，随着训练的进行，两个模型的 mAP50 均快速提升，但 LightRT-DETR 的提升速度明显快于 RT-DETR，并且在整个训练周期内始终保持更高的 mAP50，最终达到更优的检测精度。这一显著差异不仅体现了 LightRT-DETR 在训练过程中的持续优势，也表明其在最终模型精度和泛化能力方面更具竞争力。

表 7　各模型在不同目标尺寸下的精度对比

Table 7　Comparison of detection accuracy of each model for different target sizes

Model	EfficientFastGLU-18	CGA encoder	ES-FPN	mAP50/% (large target)	mAP50-95/% (large target)	mAP50/% (small target)	mAP50-95/% (small target)
M1	×	×	×	96.8	52.3	96.0	47.5
M2	√	×	×	97.6	53.9	96.8	51.8
M3	×	√	×	98.4	54.6	97.5	52.3
M4	×	×	√	98.0	53.0	97.2	52.2
M5	√	×	√	98.3	54.7	97.5	**52.9**
M6	√	√	×	97.3	54.3	96.9	52.0
M7	×	√	√	98.4	**56.0**	97.8	52.6
M8	√	√	√	**98.5**	55.6	**98.2**	52.4

注：×为不添加模块；√为添加模块；加粗数据为最优值。

表 8 对比实验结果

Table 8 Comparative experimental results

Model	mAP50/%	mAP50-95/%	GFLOPs	Params/M
EP-RTDETR[23]	96.7	46.2	—	17.7
Improved Faster R-CNN[24]	92.9	—	—	—
LIDD-YOLO[25]	93.3	—	7.1	2.6
WiseYOLOv7[26]	96.1	57.9	136.8	57
FPFM[27]	78.86	—	—	—
LDD-Net[28]	95.90	—	—	5.13
ExtendedFPN[29]	96.20	—	—	137.00
TD-Net[30]	96.20	—	—	7.06
YOLOv5l	94.0	48.5	134.7	53.1
YOLOv8m	91.3	46.6	78.7	25.8
YOLOv8l	94.7	49.3	164.8	43.6
YOLOv9m	93.4	47.8	76.5	20.0
YOLOv9e	95.3	51.1	189.2	57.3
LightRT-DETR	**98.3**	**52.9**	**37.2**	**12.8**

注：加粗数据为本文改进策略。

4.4.8 可视化分析

如图 13 所示，根据 LightRT-DETR 和 RT-DETR-R18 在 PCB 缺陷检测中的表现，可以观察到 LightRT-DETR 在多个方面展现出明显优势。对于漏孔和开路类缺陷，虽然两种模型都能成功检测，但 LightRT-DETR 表现出更高的置信度。这一特性在实际应用中尤为重要，因为高置信度的检测结果可以减少人工复查的需求，提高整体检测效率。在处理短路、杂铜和杂散等类型的缺陷时，LightRT-DETR 的优势更为明显。RT-DETR-R18 在这些缺陷类型上存在一定程度的漏检情况，而改进后的 LightRT-DETR 则展现出全面而准确的检测能力，几乎不遗漏任何缺陷。这种全面的检测能力对于保证 PCB 生产质量至关重要。此外，对于鼠咬这类较为特殊和难以识别的缺陷，RT-DETR-R18 表现出明显的不足，识别能力较弱。相比之下，经过改进的 LightRT-DETR 在这一问题上取得了显著进步，能够较好地识别和定位鼠咬缺陷。这一改进不仅展示了 LightRT-DETR 的技术优势，也凸显了其在处理复杂和多样化 PCB 缺陷方面的潜力。

图 14 展示了原始标注、RT-DETR-R18 和 LightRT-

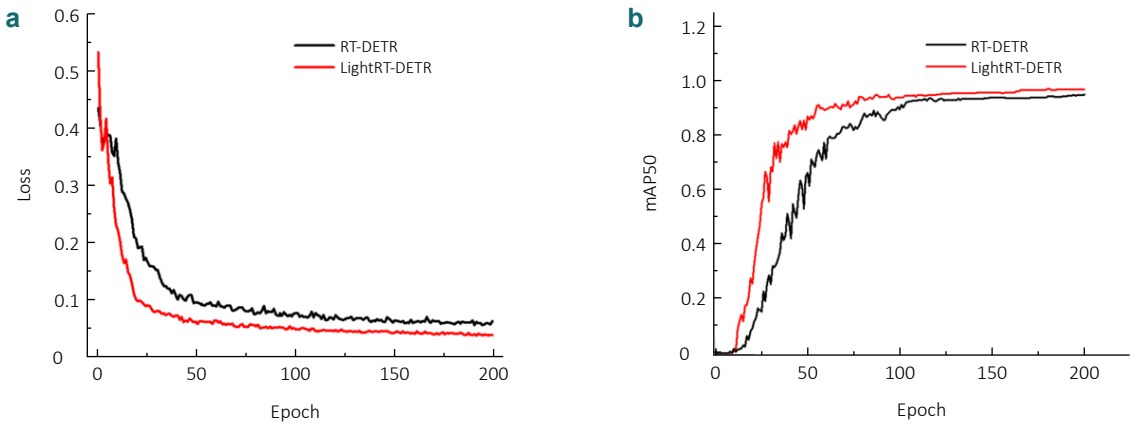


图 12 损失值图像对比。(a) 损失值随 epoch 变化曲线；(b) mAP50 随 epoch 变化曲线

Fig. 12 Comparison of loss values. (a) Curves of loss value versus epoch; (b) Curves of mAP50 versus epoch

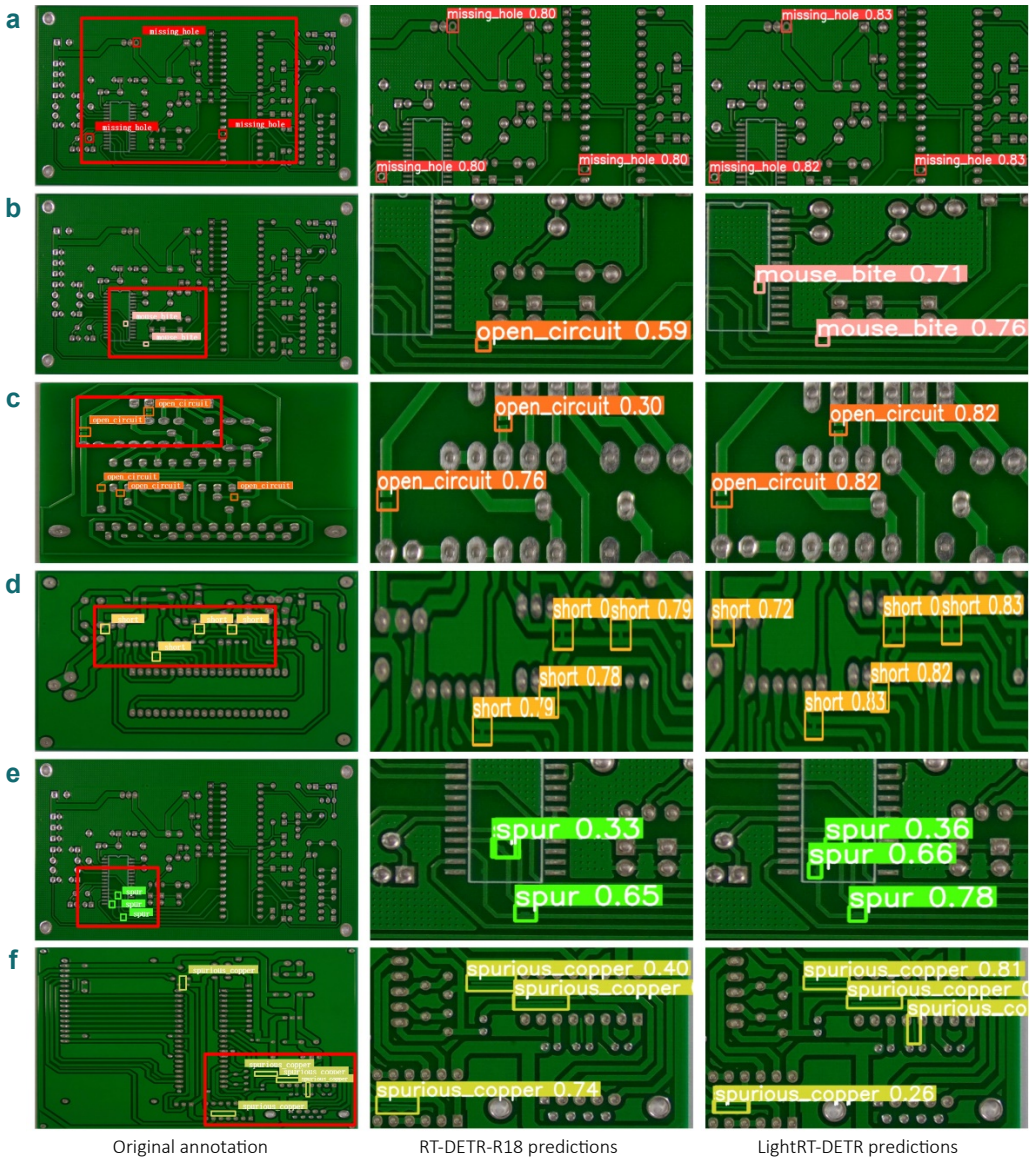


图 13 改进前后检测结果对比。(a~f) 分别代表漏孔、鼠咬、开路、短路、杂散和杂铜的原始标注及模型推理效果

Fig. 13 Comparison of detection results before and after improvement. (a-f) Represent the original annotations and model predictions results for missing hole, mouse bite, open circuit, short circuit, spur, and spurious copper, respectively

DETR 在 PCB 缺陷检测中的热力图对比。分析结果表明，LightRT-DETR 相较于 RT-DETR-R18 取得了显著进步。LightRT-DETR 生成的热力图呈现出更加集中和明确的热点分布，特别是在 PCB 板的关键区域，如 PCB 开路缺陷处。这种改进表明 LightRT-DETR 具有更高的定位精度和更强的特征提取能力，能够更准确地识别潜在的缺陷位置。相比之下，RT-DETR-R18 的热点分布相对分散，显示出在具体缺陷定位方面的不确定性。LightRT-DETR 的这一优化不仅提高了检测的准确性，还降低了误检率，对 PCB 制造质量控制的提升具有重要意义。

5 结 论

本文针对 PCB 缺陷检测中面临的挑战，提出了基于 RT-DETR 的 LightRT-DETR 算法优化。通过设计高效主干网络 EfficientFastGLU-18、引入级联分组注意力机制、提出高效选择性特征金字塔网络等一系列措施，显著提升了模型对各类 PCB 缺陷的检测能力和精度。实验结果表明，LightRT-DETR 在检测精度、速度和计算效率方面均显示出显著优势。未来，将继续深化对模型性能的研究，进一步优化 PCB 缺陷检测技术，为推

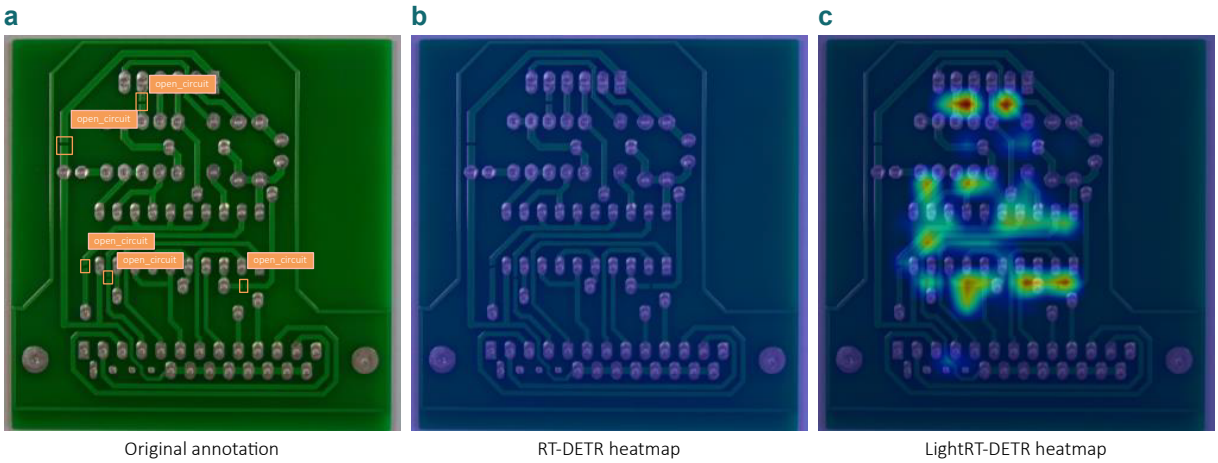


图 14 改进前后热力图对比。(a) 原始标注图；(b) RT-DETR 热力图；(c) LightRT-DETR 热力图

Fig. 14 Comparison of heatmaps before and after improvement. (a) Original annotation image; (b) RT-DETR heatmap; (c) LightRT-DETR heatmap

动智能制造技术的发展做出贡献

作者贡献

希治远：研究设计、论文初稿撰写、算法实现与验证；梅婷：参与实验设计与结果分析；林珊玲：负责研究方向把控、论文结构审阅与最终定稿；林志贤，林坚普，吕珊红：协助完成基准对比实验；郭太良：提供实验理论指导。

利益冲突

所有作者声明无利益冲突

参考文献

1. Lan Z, Hong Y, Li Y. An improved YOLOv3 method for PCB surface defect detection[C]//*2021 IEEE International Conference on Power Electronics, Computer Applications (ICPECA)*, 2021: 1009–1015. https://doi.org/10.1109/ICPECA51329.2021.9362675.
2. 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*, 2016: 779–788. https://doi.org/10.1109/CVPR.2016.91.
3. 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*, 2017: 2999–3007. https://doi.org/10.1109/ICCV.2017.324.
4. Liu W, Anguelov D, Erhan D, et al. SSD: single shot MultiBox detector[C]//*14th European Conference on Computer Vision*, 2016: 21–37. https://doi.org/10.1007/978-3-319-46448-0_2.
5. 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*, 2014: 580–587. https://doi.org/10.1109/CVPR.2014.81.
6. Girshick R. Fast R-CNN[C]//*Proceedings of the 2015 IEEE International Conference on Computer Vision*, 2015: 1440–1448. https://doi.org/10.1109/ICCV.2015.169.
7. Ren S Q, He K M, Girshick R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[C]//*Proceedings of the 29th International Conference on Neural Information Processing Systems*, 2015: 91–99.
8. He K M, Gkioxari G, Dollár P, et al. Mask R-CNN[C]//*Proceedings of the 2017 IEEE International Conference on Computer Vision*, 2017: 2980–2988. https://doi.org/10.1109/ICCV.2017.322.
9. Lian J, Wang L T, Liu T Y, et al. Automatic visual inspection for printed circuit board via novel Mask R-CNN in smart city applications[J]. *Sustain Energy Technol Assess*, 2021, **44**: 101032.
10. Li Y T, Kuo P, Guo J I. Automatic industry PCB board DIP process defect detection with deep ensemble method[C]//*2020 IEEE 29th International Symposium on Industrial Electronics (ISIE)*, 2020: 453–459. https://doi.org/10.1109/ISIE45063.2020.9152533.
11. Xin H J, Chen Z B, Wang B Y. PCB electronic component defect detection method based on improved YOLOv4 algorithm[J]. *J Phys Conf Ser*, 2021, **1827**(1): 012167.
12. Xie H F, Li Y, Li X C, et al. A method for surface defect detection of printed circuit board based on improved YOLOv4[C]//*2021 IEEE 2nd International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE)*, 2021: 851–857. https://doi.org/10.1109/ICBAIE52039.2021.9390006.
13. Wu L G, Zhang L, Zhou Q. Printed circuit board quality detection method integrating lightweight network and dual attention mechanism[J]. *IEEE Access*, 2022, **10**: 87617−87629.
14. Wang X, Gao J S, Hou B J, et al. A lightweight modified YOLOX network using coordinate attention mechanism for PCB surface defect detection[J]. *IEEE Sens J*, 2022, **22**(21): 20910−20920.
15. Tao Z Y, He Y, Lin S, et al. Surface defect detection of solar cells using local and global feature fusion[J]. *Opto-Electron Eng*, 2024, **51**(1): 230292.
陶志勇, 何燕, 林森, 等. 局部和全局特征融合的太阳能电池片表面缺陷检测[J]. 光电工程, 2024, **51**(1): 230292.
16. Carion N, Massa F, Synnaeve G, et al. End-to-end object detection with transformers[C]//*16th European Conference on Computer Vision*, 2020: 213–229.

https://doi.org/10.1007/978-3-030-58452-8_13.
17. Zhao Y, 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*, 2024: 16965–16974. https://doi.org/10.1109/CVPR52733.2024.01605.
18. He K M, Zhang X Y, Ren A Q, et al. Deep residual learning for image recognition[C]//*Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition*, 2016: 770–778. https://doi.org/10.1109/CVPR.2016.90.
19. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//*Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017: 6000–6010.
20. Liu X Y, Peng H W, Zheng N X, et al. EfficientViT: memory efficient vision transformer with cascaded group attention[C]//*Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023: 14420–14430. https://doi.org/10.1109/CVPR52729.2023.01386.
21. Hu J, Shen L, Sun G. Squeeze-and-excitation networks[C]//*Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018: 7132–7141. https://doi.org/10.1109/CVPR.2018.00745.
22. Fan Q H, Huang H B, Chen M R, et al. RMT: retentive networks meet vision transformers[C]//*Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024: 5641–5651. https://doi.org/10.1109/CVPR52733.2024.00539.
23. Wu B B, Zhang L H, Liu J W, et al. Improved EP-RTDETR based surface defect detection on PCB[J]. *Manuf Technol Mach Tool*, 2025, (3): 139−148.
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.
24. Chen R X, Zhan Z, Hu X L, et al. Printed circuit board defect detection based on the multi-attentive faster RCNN under noise interference[J]. *Chin J Sci Instrum*, 2021, **42**(12): 167−174.
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.
25. Luo S, Xu Y P, Zhang C L, et al. LIDD-YOLO: a lightweight industrial defect detection network[J]. *Meas Sci Technol*, 2024, **36**(1): 0161b5.
26. Chen J H, Xu J, Wang Z F. PCB defect detection based on deep learning[J]. *Mod Electron Tech*, 2025, **48**(8): 7−12.
陈建豪, 徐洁, 汪志锋. 基于深度学习的 PCB 缺陷检测[J]. 现代电子技术, 2025, **48**(8): 7−12.
27. Wang H D, Xie J, Xu X Y, et al. Few-shot PCB surface defect detection based on feature enhancement and multi-scale fusion[J]. *IEEE Access*, 2022, **10**: 129911−129924.
28. Zhang L X, Chen J S, Chen J G, et al. LDD-Net: lightweight printed circuit board defect detection network fusing multi-scale features[J]. *Eng Appl Artif Intell*, 2024, **129**: 107628.
29. Li C J, Qu Z, Wang S Y, et al. A method of defect detection for focal hard samples PCB based on extended FPN model[J]. *IEEE Trans Compon Packaging Manuf Technol*, 2022, **12**(2): 217−227.
30. Shao R, Zhou M L, Li M, et al. TD-Net: tiny defect detection network for industrial products[J]. *Complex Intell Syst*, 2024, **10**(3): 3943−3954.

作者简介



希治远 (2000-)，男，硕士研究生，主要研究方向为工业缺陷检测。

E-mail：xzy_20001222@163.com



【通信作者】林珊玲 (1991-)，女，福建泉州人，博士，讲师，硕士研究生导师，福州大学先进制造学院电子信息系教师，主要从事主要研究新型显示技术、平板显示器驱动系统图像处理技术等。

E-mail：sllin@fzu.edu.cn

Printed circuit board defect detection based on LightRT-DETR algorithm

Xi Zhiyuan[1,3], Mei Ting[1,3], Lin Shanling[2,3*], Lin Zhixian[1,2,3], Lin Jianpu[2,3], Lü Shanhong[2,3], Guo Tailiang[1,3]

[1] School of Physics and Information Engineering, Fuzhou University, Fuzhou, Fujian 350116, China;
[2] School of Advanced Manufacturing, Fuzhou University, Quanzhou, Fujian 362000, China;
[3] Fujian Optoelectronic Information Science and Technology Innovation Laboratory, Fuzhou, Fujian 350116, China

Abstract

Objective The manufacturing of printed circuit boards (PCBs) constitutes the foundational infrastructure of the contemporary electronics

Foundation item: National Key Research and Development Program of China (2021YFB3600603)
*Correspondence: Lin S L, E-mail: sllin@fzu.edu.cn

industry. As the sector evolves in response to consumer demands for increasingly miniaturized, sophisticated, and high-performance devices, the physical characteristics of PCBs have undergone significant transformation. Modern boards are characterized by high-density interconnects, multi-layered architectures, and microscopic component placement. In this context, the role of quality control has shifted from a supplementary process to a critical determinant of product viability and yield. Traditionally, defect detection has relied on manual inspection or standard automated optical inspection (AOI) systems. However, these conventional methods are increasingly challenged by the escalating complexity of modern PCB designs. Manual inspection is labor-intensive, subjective, and prone to fatigue-induced errors, rendering it unsuitable for mass production. Similarly, traditional AOI systems based on pattern matching or simple image processing often struggle with the subtle variations and complex backgrounds found in high-density PCBs. Consequently, there is a growing necessity for more intelligent and adaptive inspection solutions capable of discerning defects with high precision. Current state-of-the-art detection methods, particularly those leveraging deep learning paradigms, encounter a distinct set of challenges when applied to the specific constraints of industrial manufacturing environments. The first challenge is the significant computational burden associated with high-precision models. Accurately detecting minute defects typically necessitates high-resolution imagery; however, processing such data with deep convolutional neural networks (CNNs) or complex Transformer models consumes substantial computational resources and memory. This computational demand often translates to prohibitive hardware costs for widespread deployment on factory floors. The second challenge is scale variation, which presents a persistent hurdle. PCB defects manifest in widely varying sizes, from tiny pinholes barely visible to the naked eye to larger structural errors like open circuits. A detection model must possess the multi-scale capability to identify these diverse anomalies with uniform accuracy. The third challenge is the imperative for real-time performance. In high-velocity production lines, inspection algorithms must operate at speeds commensurate with manufacturing throughput to avoid creating bottlenecks. Unfortunately, many existing high-accuracy models are computationally too intensive for real-time inference, forcing manufacturers to make an undesirable trade-off between inspection speed and detection precision. These limitations highlight the need for an optimized architecture that balances resource efficiency with robust detection capabilities.

Methods To address these pressing challenges effectively, this study proposes LightRT-DETR, a lightweight real-time detection transformer specifically optimized for PCB defect detection. Building upon the architectural principles of the real-time detection Transformer (RT-DETR), our approach introduces a series of strategic modifications designed to mitigate computational costs significantly while maintaining, and in some aspects enhancing, detection accuracy. The design of LightRT-DETR is characterized by three primary technical innovations: the implementation of the EfficientFastGLU-18 backbone, the integration of a cascaded group attention (CGA) mechanism, and the development of an efficient selective feature pyramid network (ES-FPN). These components work in concert to refine the model's ability to extract relevant features from complex backgrounds without incurring the heavy computational penalties associated with standard transformer architectures. The overarching goal is to create a model that is both agile enough for edge deployment and accurate enough to meet rigorous quality standards. To resolve the inefficiency of traditional backbones like ResNet-18 when deployed on edge computing devices, we propose the integration of the EfficientFastGLU-18 backbone. This novel structure synthesizes partial convolution (PConv) with convolutional gated linear units (GLU). PConv operates by applying convolution to only a subset of input channels while leaving the remainder untouched for the next layer. This strategy effectively reduces memory access and floating-point operations (FLOPs) without compromising the flow of information. By coupling this efficient convolution strategy with GLU, which provides robust non-linear transformation capabilities, the backbone maintains strong feature extraction abilities. This is particularly crucial for identifying defects against the complex, texture-rich backgrounds of PCBs, where differentiating between a valid trace and a defect requires discerning subtle feature variations. Furthermore, to address the computational expense of standard multi-head attention mechanisms, we introduce a cascaded group attention (CGA) mechanism. Instead of processing all feature channels simultaneously, CGA splits feature maps into distinct groups and processes them through a cascaded structure. This approach allows the network to refine features progressively, as the output of one group's attention computation can inform the next, creating a hierarchical refinement process. This mechanism appears particularly beneficial for defects possessing intricate or fine-grained details, such as "mouse bites" or "spurs", which are frequently overlooked by coarser, global attention mechanisms. Finally, we tackle multi-scale defect detection through the efficient selective feature pyramid network (ES-FPN). This redesigned module employs selective feature integration and incorporates RetBlockC3 modules enhanced with manhattan self-attention (MaSA). Unlike standard self-attention, MaSA is specifically designed to capture long-range dependencies in horizontal and vertical directions. Given that PCB layouts consist largely of orthogonal lines, this domain-specific adaptation enables more effective recognition of defects across widely varying scales by aligning the attention mechanism with the geometric priors of the PCB domain.

Results and Discussions The efficacy of LightRT-DETR was validated through extensive experiments on a public PCB defect dataset, revealing a significant improvement in both efficiency and accuracy compared to current state-of-the-art models, including the widely utilized YOLO series. In terms of quantitative performance, LightRT-DETR achieved a mean average precision (mAP50) of 98.3% and a

stricter metric, mAP50-95, of 52.9%. When benchmarked against the advanced YOLOv9e model, our approach demonstrated an improvement of 3.0% in mAP50 and 1.8% in mAP50-95. These metrics suggest that the model is not only capable of detecting defects but also does so with high localization accuracy, minimizing false positives and false negatives—a critical requirement for industrial quality control. The experimental setup ensured fair comparisons by utilizing consistent hyperparameters and training environments, further reinforcing the validity of the observed performance gains.

Conclusions Perhaps the most distinct advantage of LightRT-DETR lies in its computational efficiency. The model requires approximately 37.2 GFLOPs, representing an 80.3% reduction in computational cost compared to the baseline RT-DETR model. Similarly, the parameter count is reduced to 12.8 million, a 77.7% decrease. This substantial reduction has practical implications: it facilitates deployment on cost-effective hardware with limited computing power or allows the system to operate at significantly higher frame rates on standard hardware.

Keywords: defect detection; printed circuit board; backbone network; feature pyramid network; computational redundancy

Xi Z Y, Mei T, Lin S L et al. Printed circuit board defect detection based on LightRT-DETR algorithm[J]. *Opto-Electron Eng*, 2026, **53**(2): 250117; DOI: 10.12086/oee.2026.250117

Opto-Electronic Journals Group
www.oejournal.org

扫描二维码，获取 PDF 全文