Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

DGFN-DETR: Lightweight Dual-Stream Gated Transformer for Real-Time PCB Defect Detection

huan yang
Shanghai University of Engineering Sciences
hua han
Shanghai University of Engineering Sciences
zhenrui yang
Shanghai University of Engineering Sciences
yuanyuan li
liyuanyuan@sues.edu.cn

Shanghai University of Engineering Sciences

Research Article

Keywords: PCB defect detection, Lightweight network, Transformer, Real-time inference, Model compression, Multi-scale feature fusion

Posted Date: March 26th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9129454/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

DGFN-DETR: Lightweight Dual-Stream Gated Transformer for Real-Time PCB Defect Detection

Huan Yang · Hua Han · Zhenrui Yang · Yuanyuan Li

Received: date / Accepted: date

Abstract Surface defects on printed circuit boards significantly affect product quality and device reliability. Lightweight and efficient PCB defect detection methods are essential for reducing dependence on manual inspection and enabling deployment on resource-constrained industrial systems. Existing approaches struggle to achieve an optimal balance between model complexity and detection accuracy, particularly when identifying small defects embedded within complex circuit patterns. In this paper, we propose DGFN-DETR (Dual-Stream Gated Feature Network Detection Transformer), a lightweight detection framework specifically designed to address PCB defect detection challenges with minimal computational overhead. Our method introduces several key innovations for model compression and efficiency: a novel Dual-Stream Gated Backbone (DSG) that adaptively fuses lightweight and enhanced pathways for efficient multi-scale feature extraction; an enhanced encoder incorporating polarized linear attention and spatially-enhanced feedforward networks that achieve linear complexity; dilated reparameterized convolution blocks that enable multi-scale receptive fields without additional inference cost through structural reparameterization; weighted stride downsampling that preserves fine-grained features while reducing computational burden; and an improved loss function for bounding box regression of irregular defect geometries. Experiments on the PKU-PCB dataset demonstrate that our method achieves 97.3% mAP50 with only 10.2M parameters and 19.4 GFLOPs computational cost, representing a 48.7% reduction in parameters and 66.0% reduction in GFLOPs compared with the baseline RT-DETR-R18. The lightweight design yields improved inference speed of 112 FPS with 8.9 ms average latency and 11.2 ms 95th percentile latency, meeting real-time requirements for production line deployment. Comprehensive evaluation including memory footprint (39.2 MB) and power consumption (68W) validates practical deployability on edge devices. Cross-dataset evaluation on DeepPCB further validates the generalization capability of the proposed lightweight architecture.

Keywords PCB defect detection · Lightweight network · Transformer · Real-time inference · Model compression · Multi-scale feature fusion

1 Introduction

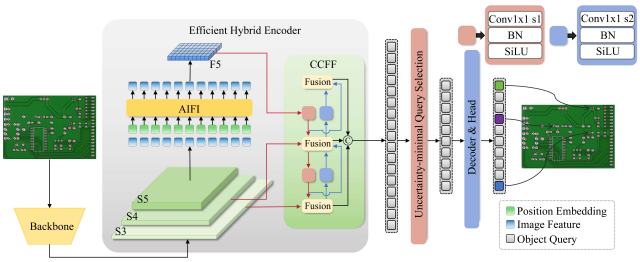


Fig. 1 RT-DETR baseline network architecture.

Yuanyuan Li (corresponding author) · Huan Yang · Hua Han · Zhenrui Yang
School of Electronic and Electrical Engineering, Shanghai University of Engineering Science, Guangfulin Road, Songjiang District, Shanghai 201620, China
E-mail: liyuanyuan@sues.edu.cn (Yuanyuan Li)
E-mail: M325124248@sues.edu.cn (Huan Yang)
E-mail: 2070967@mail.dhu.edu.cn (Hua Han)
E-mail: M325124244@sues.edu.cn (Zhenrui Yang)

Industrial defect detection plays a critical role in manufacturing quality control, requiring both high accuracy and computational efficiency. Printed circuit boards (PCBs) are fundamental components in electronic devices and are susceptible to various defects including missing holes, mouse bites, open circuits, shorts, spurious copper, and spurs. If undetected, these defects can lead to unreliable device operation and system failures. Manual inspection is limited by subjective judgment, operator fatigue, and inconsistent detection rates. With deep learning revolutionizing machine vision through automatic feature extraction and end-to-end learning paradigms, PCB defect detection has attracted significant research attention.

Modern object detection architectures have evolved from two-stage detectors such as Faster R-CNN [1] to one-stage detectors like the YOLO family [2]. For PCB defect detection, various deep learning approaches have been proposed. Ding et al. [3] developed TDD-net, a lightweight network combining channel pruning with knowledge distillation for efficient detection. Liu et al. [4] proposed an automatic optical inspection system integrating image processing with convolutional neural networks for solder joint defects. Hu et al. [5] introduced a hybrid method combining traditional image processing with deep learning for surface-mount device inspection. Tang et al. [6] presented a deep learning framework with adaptive feature pyramid networks addressing multi-scale PCB defect detection challenges. Wang et al. [7] developed a lightweight detection network utilizing depthwise separable convolutions for real-time performance. Chen et al. [8] proposed a multi-scale feature fusion approach for detecting defects of varying sizes on PCB surfaces. Li et al. [9] introduced an attention mechanism combining spatial and channel information for improved feature extraction. Yang et al. [10] presented an advanced deep learning framework with adaptive feature aggregation strategies for handling complex PCB patterns. However, these methods often employ complex architectures with high parameter counts, limiting their practical deployment on edge devices.

The introduction of Transformer-based detectors, exemplified by DETR [11], represents a paradigm shift in object detection. Unlike conventional methods requiring complex hand-crafted components such as NMS and anchor generation, DETR leverages attention mechanisms for end-to-end detection. However, DETR suffers from high computational cost and slow convergence, limiting its practical deployment. RT-DETR [12] addresses these limitations through hybrid encoding strategies that achieve real-time performance while maintaining end-to-end detection capabilities. The RT-DETR architecture is shown in Fig. 1. Kim et al. [13] proposed a lightweight Transformer architecture optimized for edge deployment. Recent advances in lightweight real-time detection have demonstrated promising results: Wen et al. [34] integrated Swin Transformer blocks into YOLOv5s for real-time coal-gangue detection, Xiang et al. [38] proposed FRT-DETR for industrial surface defect detection with reduced computational complexity, and Li et al. [36] introduced GSConv-based lightweight designs achieving state-of-the-art accuracy-efficiency trade-offs. Despite improvements in detection accuracy, these methods still face challenges in achieving optimal accuracy-efficiency balance, particularly regarding model size and computational cost for edge deployment.

Despite these advances, existing methods exhibit several limitations from a lightweight design perspective: insufficient balance between model complexity and detection accuracy, inadequate consideration of parameter efficiency for edge deployment, lack of effective multi-scale feature fusion mechanisms with low computational overhead, and insufficient handling of irregular defect geometries in loss functions. To address these challenges, we propose DGFN-DETR, a lightweight yet accurate framework built upon RT-DETR with architectural innovations tailored for efficient PCB defect detection. The main contributions are:

- A novel Dual-Stream Gated Backbone (DSG) that adaptively fuses lightweight and enhanced pathways, achieving 41.2% parameter reduction and 47.7% computational cost reduction compared to RT-DETR-R18 backbone while maintaining competitive accuracy.
- An enhanced encoder incorporating Polarized Linear Attention with reduced complexity from $O(N^2)$ to effectively linear in sequence length, Spatially-enhanced Feedforward Networks operating in frequency domain for efficient global feature aggregation, and novel normalization layers for improved feature representation.
- Dilated reparameterized convolution blocks that enable multi-scale receptive fields during training while collapsing to single convolutions during inference, eliminating multi-branch overhead, and weighted stride downsampling modules that preserve fine-grained information with minimal computational cost.
- A composite WFS-IoU loss function that effectively handles irregular geometries and addresses class imbalance, enabling improved detection of small and deformable defects.

– Comprehensive real-time validation including latency analysis (8.9 ms average, 11.2 ms 95th percentile), memory profiling (39.2 MB model size), and power measurement (68W), demonstrating practical deployability on edge platforms.

The resulting lightweight architecture achieves 97.3% mAP50 with only 10.2M parameters (48.7% reduction) and 19.4 GFLOPs (66.0% reduction), while also delivering improved inference speed of 112 FPS, demonstrating that lightweight design and real-time performance can be achieved simultaneously without sacrificing accuracy.

2 Methods

2.1 Overall Architecture

The DGFN-DETR model is designed to address key challenges in industrial PCB defect detection with minimal computational overhead. As illustrated in Fig. 2, our network comprises several specialized components that tackle different defect categories, capture subtle visual differences within complex circuit patterns, and handle irregular defect shapes while maintaining a lightweight architecture. For a given PCB image, the DSG backbone extracts hierarchical features through two complementary pathways, precisely capturing both fine details and overall layout structure with reduced parameter count. The extracted multi-scale features (P3, P4, P5) are processed through our PME-Layer (Polarized-Mona-Enhanced Layer) for enhanced representation with linear complexity. The feature pyramid, augmented with dilated reparameterized blocks, enables multi-scale processing without inference-time overhead through structural reparameterization, while the RT-DETR decoder generates accurate predictions. This lightweight design ensures efficient deployment on resource-constrained edge devices while maintaining high detection accuracy.

2.2 Dual-Stream Gated Backbone

2.2.1 Architecture Design

PCB defect detection requires capturing both small defects (e.g., hairline cracks, solder bridges) and their contextual relationships within complex circuit layouts. The DSG backbone addresses this challenge through two complementary pathways that are dynamically fused via multi-scale adaptive spatial attention gates. The overall DSG architecture is illustrated in Fig. 2, with detailed gating mechanisms shown in Fig. 3. The architecture begins with a shared HGStem (Hierarchical Gradient Stem) module for initial feature extraction, generating 64-channel feature maps at P2 level (1/4 resolution). This shared representation captures fundamental circuit patterns before branching into specialized streams, reducing redundant parameters and computations critical for lightweight design.

The lightweight stream employs C2f modules [14] optimized for computational efficiency while maintaining sensitivity to small defects. The enhanced stream utilizes HGBlock series with hierarchical gradient mechanisms to extract complex defect characteristics. The dual-stream design enables effective feature extraction with significantly fewer parameters than single heavy-weight backbones. The complete feature extraction at each stage i is formulated as:

$$\mathbf{F}_l^i = \text{C2f}(\text{Conv}(\mathbf{F}_l^{i-1}, s=2)), \tag{1}$$

where $i \in \{2,3,4\}$, and

$$\begin{aligned}\mathbf{F}_e^i = \text{HGBlock}^{n_i}(\text{DWConv}(\mathbf{F}_e^{i-1}, s=2)),\\ n_i \in \{3,9,1\},\end{aligned} \tag{2}$$

where $\mathbf{F}_l^i$ and $\mathbf{F}_e^i$ denote features at stage i for lightweight and enhanced streams respectively.

2.2.2 Multi-Scale Adaptive Spatial Attention Gate

PCB defects exhibit diverse characteristics across different scales. The Multi-Scale Adaptive Spatial Attention Gate (MSGA) adaptively modulates the contributions from each stream based on local defect patterns, following the MSA^{2}Net architecture [15]. The gating mechanism operates with minimal parameter overhead, maintaining the lightweight nature of the overall architecture.

Multi-scale fusion aggregates contextual information using multiple pooling operations. The fused features are computed as:

$$\begin{aligned}\mathbf{F}_{\text{fused}} =& \text{PWConv}(\text{Concat}[\text{DW}(\mathbf{F}_e),\\ &\text{DW-D}(\mathbf{F}_e), \text{MaxPool}(\mathbf{F}_l),\\ &\text{AvgPool}(\mathbf{F}_l)]),\end{aligned} \tag{3}$$

where PWConv denotes point-wise convolution, DW represents depthwise convolution, DW-D is dilated depthwise convolution, and Concat denotes concatenation.

Spatial attention weights are computed through channel-wise softmax to emphasize defect-relevant regions:

$$\mathbf{W}_{\text{spatial}} = \text{Softmax}_{\text{ch}}(\text{PWConv}(\mathbf{F}_{\text{fused}})), \tag{4}$$

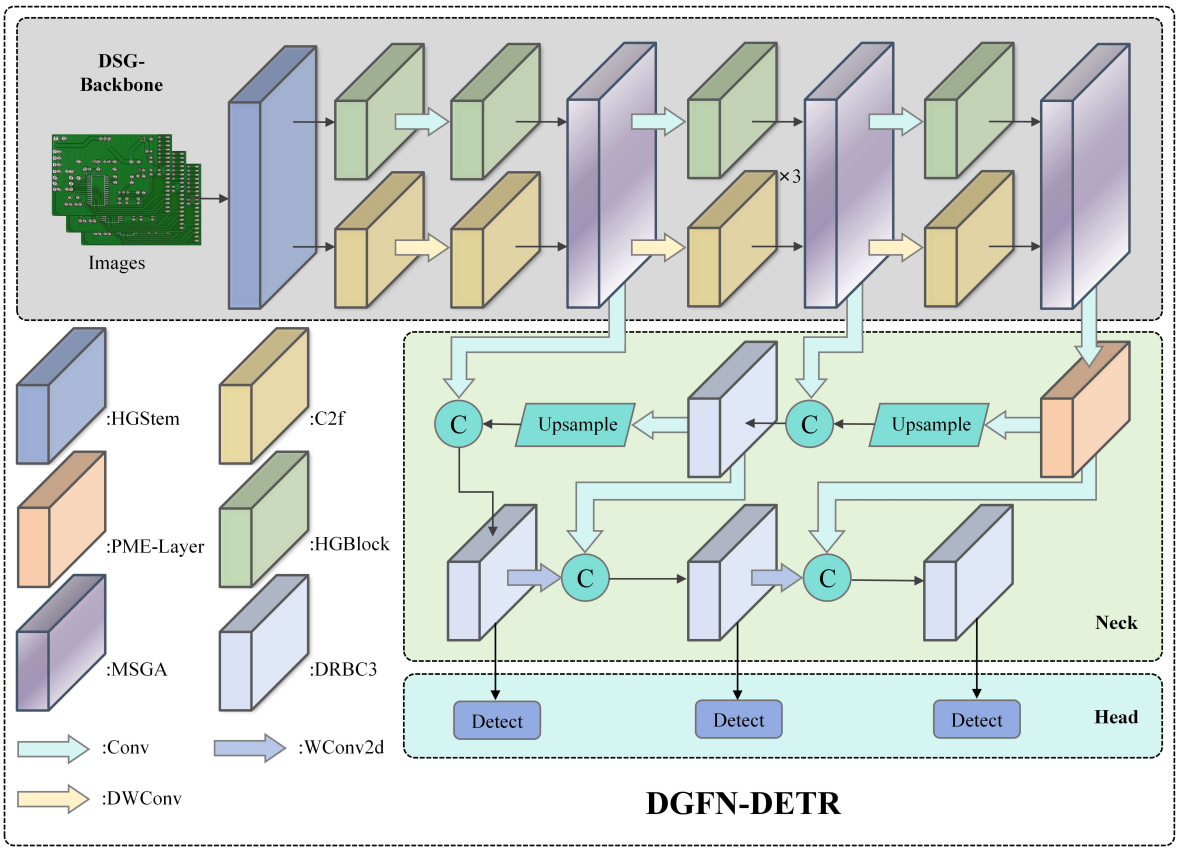


Fig. 2 DGFN-DETR framework architecture.

where $\mathbf{W}_{\text{spatial}} \in \mathbb{R}^{2\times H\times W}$ contains attention weights for both enhanced and lightweight streams.

The final output integrates spatial selection, cross-modulation, and gated fusion mechanisms:

$$\begin{aligned}\mathbf{F}_{\text{out}} =&\text{PWConv}(\sigma(\text{PWConv}(\\ &\mathbf{F}''_e \odot \mathbf{F}''_l)) \odot \mathbf{F}_e),\end{aligned} \tag{5}$$

where σ denotes sigmoid activation, $\odot$ represents element-wise multiplication, and $\mathbf{F}''_e, \mathbf{F}''_l$ are modulated features from both streams.

2.3 PME-Layer: Enhanced Encoder Design

PCB defect detection requires simultaneous capture of global context and local detail precision. The PME-Layer integrates four complementary components, as illustrated in Fig. 4, each designed to maximize feature representation quality while minimizing computational complexity.

The Polarized Linear Attention [16] achieves reduced complexity through polarity decomposition, significantly reducing computational cost for high-resolution feature maps. Given input features $\mathbf{X}$, queries $\mathbf{q}$ and keys $\mathbf{k}$ are decomposed as $\mathbf{q} = \mathbf{q}^+ - \mathbf{q}^-$ and $\mathbf{k} = \mathbf{k}^+ - \mathbf{k}^-$ where $\mathbf{q}^+ = \max(\mathbf{q}, 0)$. For notational simplicity, we denote $\phi_t = \phi([\mathbf{q}_t^+; \mathbf{q}_t^-])$. The attention output processes dual streams concatenated together:

$$\begin{aligned}\mathbf{o}_t = \Bigg[&\phi_t \frac{\sum_{i=1}^{N} \phi([\mathbf{k}_i^+; \mathbf{k}_i^-])^\top \mathbf{v}_i^s}{\phi_t \sum_{j=1}^{N} \phi([\mathbf{k}_j^+; \mathbf{k}_j^-])^\top} \odot \mathbf{G}^s; \\ &\phi_t \frac{\sum_{i=1}^{N} \phi([\mathbf{k}_i^-; \mathbf{k}_i^+])^\top \mathbf{v}_i^o}{\phi_t \sum_{j=1}^{N} \phi([\mathbf{k}_j^-; \mathbf{k}_j^+])^\top} \odot \mathbf{G}^o\Bigg],\end{aligned} \tag{6}$$

where $\mathbf{v}^s$ and $\mathbf{v}^o$ represent value halves for same-signed and opposite-signed interactions, $\mathbf{G}^s, \mathbf{G}^o$ are learnable coefficient matrices, and $\phi(\cdot)$ is the feature mapping with learnable power rescaling. This linear attention mechanism reduces computational complexity from $O(N^2 d)$ to $O(Nd^2)$, where $d \ll N$ for typical feature dimensions, effectively achieving linear scaling in sequence length N for high-resolution inputs.

The EDFFN [17] captures spatial relationships through frequency-domain operations with Fast Fourier Transform (FFT), which can be efficiently computed with $O(N \log N)$ complexity:

$$\begin{aligned}\mathbf{h}_1 &= \text{GELU}(\text{PConv}(\text{DConv}(\mathbf{h}_{\text{in}}))), \\ \mathbf{F}_{\text{output}} &= \mathbf{h}_{\text{in}} + \text{IFFT}(\mathbf{W} \odot \text{FFT}(\mathbf{h}_1)),\end{aligned} \tag{7}$$

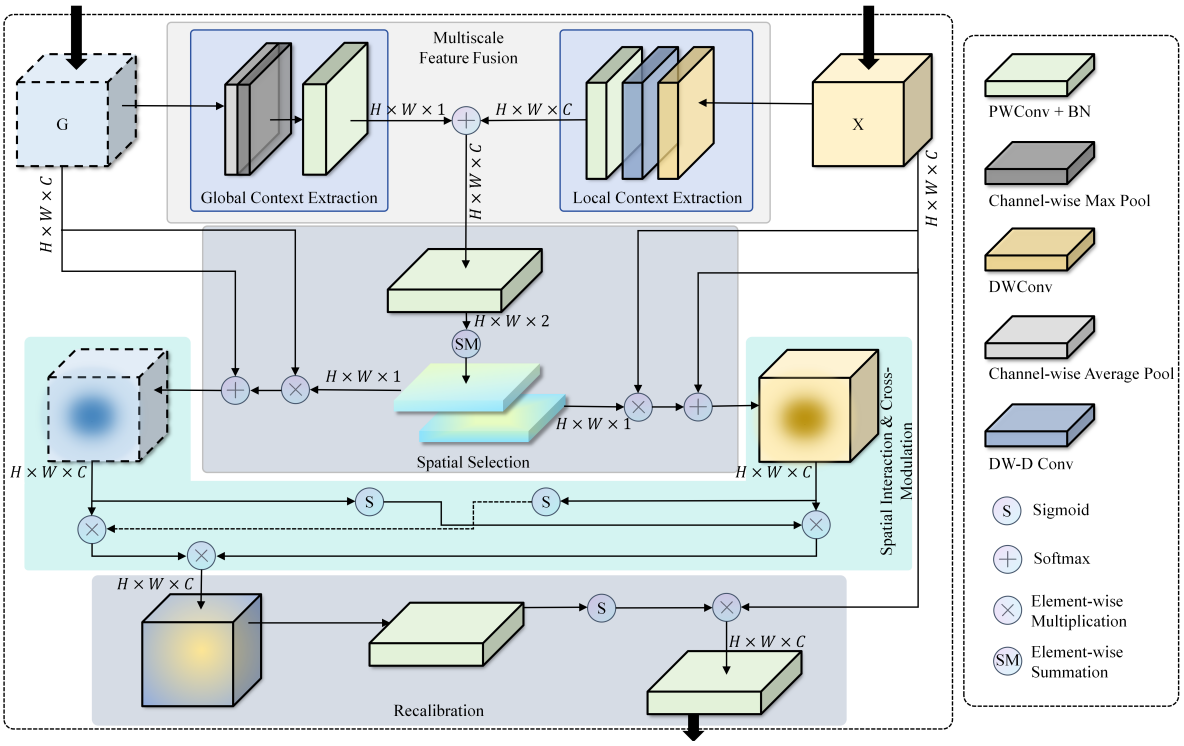


Fig. 3 Multi-Scale Adaptive Spatial Attention Gate structure.

where $\mathbf{W}$ is a learnable quantization matrix determining frequency preservation. The frequency-domain processing enables global feature aggregation with lower complexity than spatial convolutions.

The Mona [18] component enables rapid adaptation through Scaled LayerNorm:

$$\mathbf{x}_{\text{norm}} = s_1 \odot |\mathbf{x}_0|_{LN} + s_2 \odot \mathbf{x}_0, \tag{8}$$

and multi-scale depth-wise convolutions:

$$f_{dw} = \mathbf{x} + \text{avg}\left(\sum_{i=1}^{3} \boldsymbol{\omega}_{dw}^{i} \hat{\otimes} \mathbf{x}\right) \tag{9}$$

with kernels 3×3, 5×5, 7×7. The complete computation is:

$$\mathbf{x} = \mathbf{x}_0 + U_l \sigma(f_{pw}(f_{dw}(D_l(\mathbf{x}_{\text{norm}})))), \tag{10}$$

where D_l, U_l are down/up projections and σ is GELU activation.

The DynamicTanh normalization [19] replaces layer normalization with a simpler yet effective operation that reduces computational overhead:

$$\text{DyT}(\mathbf{x}) = \boldsymbol{\omega} \odot \tanh(\alpha \mathbf{x}) + \boldsymbol{\vartheta}, \tag{11}$$

where α is a learnable scalar, $\boldsymbol{\omega} \in \mathbb{R}^C$ and $\boldsymbol{\vartheta} \in \mathbb{R}^C$ are learnable per-channel scaling and shifting vectors.

The complete PME-Layer formulation integrates all components:

$$\begin{aligned}\mathbf{X}_{\text{out}} =& \text{Mona}(\text{DyT}(\mathbf{X} + \text{PolaAttn}(\mathbf{X}) \\ &+ \text{EDFFN}(\mathbf{X}))).\end{aligned} \tag{12}$$

2.4 Efficient Multi-Scale Processing Modules

To enhance multi-scale feature extraction while maintaining a lightweight architecture, we introduce two complementary modules: Dilated Reparameterized Convolution Blocks (DRBC3) and Adaptive Weighted Stride Downsampling.

DRBC3 modules, inspired by UniRepLKNet [20], enable multi-scale processing through structural reparameterization with enhanced receptive field capabilities. As shown in Fig. 5, during training, parallel branches with different dilation rates capture multi-scale features:

$$\begin{aligned}\mathbf{F}_{\text{out}} =& \text{BN}(\text{DWConv}_K(\mathbf{F}_{\text{in}})) \\ &+ \sum_{i=1}^{n} \text{BN}(\text{DWConv}_{k_i, r_i}(\mathbf{F}_{\text{in}})),\end{aligned} \tag{13}$$

where K is the target large kernel size, k_i, r_i are kernel sizes and dilation rates of parallel branches, and the equivalent kernel size is $K_{\text{equiv}} = (k-1) \times r + 1$. The key advantage for lightweight deployment is that

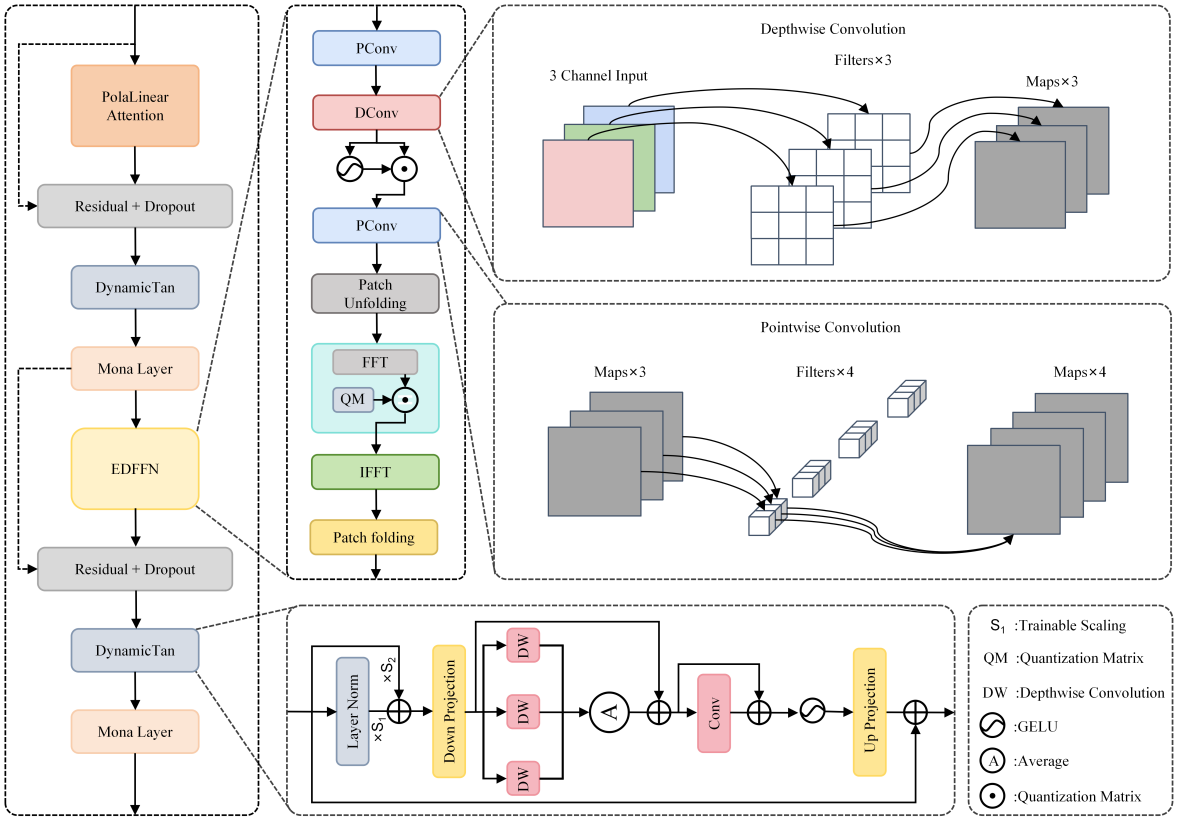


Fig. 4 PME-Layer structure.

during inference, all branches are reparameterized into a single convolution through algebraic fusion of batch normalization and convolution layers. This eliminates multi-branch computational overhead while maintaining the effective receptive field learned during training, achieving training-time expressiveness with inference-time efficiency.

For downsampling operations, we propose an adaptive weighted stride convolution module, termed wConv2d. This module employs learnable density functions [21] to adaptively adjust the sampling process, preserving small defect information that might otherwise be lost during standard stride convolutions. For a given position $\mathbf{p}$, the output feature $\mathbf{F}_{\text{out}}$ is computed as:

$$\mathbf{F}_{\text{out},f}(\mathbf{p}) = \sum_{i,j} \mathcal{N}_{i,j}^{\mathbf{p},r}(\mathbf{F}_{\text{in}}) \odot \mathbf{\Phi}_{i,j}^{r} \odot \mathbf{K}_{f,i,j}^{r}, \tag{14}$$

where $\mathcal{N}^{\mathbf{p},r}(\mathbf{F}_{\text{in}})$ is the receptive field neighborhood, $\mathbf{\Phi}^r$ is the learnable density matrix, and $\mathbf{K}_f^r$ is the convolution kernel. Together, DRBC3 and wConv2d enhance multi-scale receptive fields while preserving fine-grained features with minimal parameter overhead.

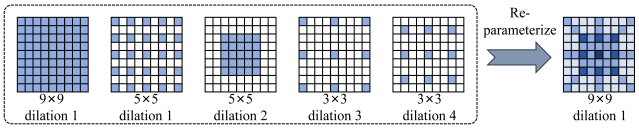


Fig. 5 DRBC3 structure and reparameterization strategy.

2.5 Loss Function

Our composite loss combines classification, localization, and geometric constraints tailored for irregular PCB defect shapes. We propose WFS-IoU, which integrates three complementary components to address critical challenges: dynamic gradient allocation for handling class imbalance, focused training on hard-to-detect samples, and geometric penalty terms for elongated defect patterns. The total loss is:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{cls}} + \lambda_1 \mathcal{L}_{\text{WFS-IoU}} + \lambda_2 \mathcal{L}_{L_1}. \tag{15}$$

The WFS-IoU loss integrates three components. First, the Wise-IoU component [22] provides dynamic gradient gain based on outlier degree:

$$r = \beta^{\delta}, \quad \beta = \frac{L_{\text{IoU}}^{*}}{\bar{L}_{\text{IoU}}}, \tag{16}$$

where L_{IoU}^{*} is the detached IoU loss of the current sample, and $\bar{L}_{\text{IoU}}$ is the mean IoU loss across the batch.

Second, the Focaler-IoU component [23] applies interval mapping to focus on challenging samples:

$$\mathrm{IoU}^{\mathrm{focaler}} = \begin{cases} 0, & \mathrm{IoU} < d \\ \frac{\mathrm{IoU}-d}{u-d}, & d \le \mathrm{IoU} \le u \\ 1, & \mathrm{IoU} > u \end{cases} \tag{17}$$

where $[d, u]$ defines the focus interval.

Third, the ShapeIoU component [24] captures geometric properties of irregular defects through width and height penalties, denoted as Ω_{shape}. The complete WFS-IoU loss integrates all three components:

$$\mathcal{L}_{\text{WFS-IoU}} = r \odot (1 - \mathrm{IoU}^{\mathrm{focaler}} + \Omega_{\mathrm{shape}}). \tag{18}$$

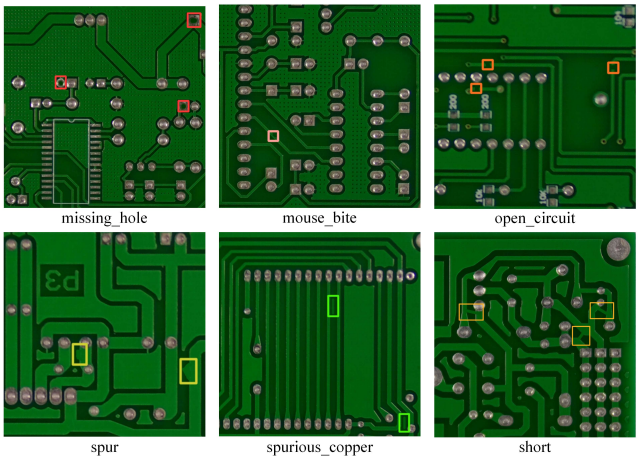


Fig. 6 PCB defect dataset samples.

3 Experiments

3.1 Experimental Setup

Experiments were conducted on the PCB defect dataset from Peking University, comprising 693 images distributed across six defect types: missing hole, mouse bite, open circuit, short, spur, and spurious copper. Following standard augmentation operations including rotation, flipping, scaling, and color jittering, the dataset was expanded to 3,481 images and randomly split into training, validation, and testing sets at an 8:1:1 ratio. Example samples are shown in Fig. 6. All models were implemented in PyTorch 2.0 and trained on NVIDIA RTX 4090 GPU. We employed the AdamW optimizer with an initial learning rate of 0.001, weight decay of 0.0001, and cosine annealing schedule. Training proceeded for 150 epochs with a batch size of 4 and input resolution of 640×640 pixels. Standard augmentation techniques included mosaic, mixup, and random affine transformations.

For comprehensive evaluation, we assess both detection accuracy and real-time performance metrics. Accuracy metrics include Precision (P), Recall (R), mAP50 and mAP in the IoU threshold range of 0.5–0.95 (mAP50-95). Efficiency metrics include model size (Parameters), computational cost (GFLOPs), and inference speed (FPS). To evaluate real-time performance, we also report average inference latency and 95th percentile latency for deterministic timing analysis, model size for deployment feasibility, and power consumption measured during sustained inference. All latency measurements exclude data loading and pre/post-processing overheads, representing pure model inference time on GPU.

3.2 Backbone Architecture Comparison

To validate the effectiveness of our proposed DSG backbone design, we performed comprehensive comparisons with several modern backbone architectures under identical experimental conditions. As shown in Table 1, we evaluated several contemporary backbones, all integrated into the same RT-DETR framework for fair comparison.

VanillaNet [29] achieves the highest mAP50 of 97.2%, but requires substantial computational resources (110.5 GFLOPs, 21.8M parameters), making it unsuitable for edge deployment where model size and computational cost are primary concerns. Lightweight alternatives such as ConvNeXtv2 [30], EfficientViT [31], and FasterNet [32] achieve reduced model complexity but sacrifice detection accuracy, reaching only 95.8%, 95.1%, and 95.3% mAP50 respectively. RepViT [33] and LSKNet [35] demonstrate balanced performance with approximately 96.0% mAP50, yet still require over 36 GFLOPs. The baseline RT-DETR-R18 achieves strong performance with 96.7% mAP50, but its computational cost of 57.0 GFLOPs and 19.9M parameters limits deployment on resource-constrained platforms. Our proposed DSG backbone achieves an optimal accuracy-efficiency trade-off, attaining 96.1% mAP50 (only 0.6 percentage points lower than RT-DETR-R18) with only 29.8 GFLOPs—a 47.7% reduction in computational cost. The parameter count is simultaneously reduced from 19.9M to 11.7M (41.2% reduction), enabling deployment on edge devices with limited memory and storage.

Table 1 Comparison of different backbone architectures on PCB defect detection.

Backbone	P (%)	R (%)	mAP50 (%)	mAP50-95 (%)	GFLOPs	Params (M)
VanillaNet [29]	97.5	96.8	97.2	56.5	110.5	21.8
ConvNeXtv2 [30]	95.5	96.8	95.8	52.2	32.2	12.4
EfficientViT [31]	94.8	95.2	95.1	51.5	27.6	10.8
FasterNet [32]	95.0	95.4	95.3	51.8	28.8	10.9
RepViT [33]	96.6	95.9	95.9	52.5	36.7	12.7
LSKNet [35]	95.7	96.0	96.0	52.6	37.9	12.7
RT-DETR-R18 [12]	95.9	96.0	96.7	53.8	57.0	19.9
DSG (Ours)	95.8	96.1	96.1	52.8	29.8	11.7

3.3 Loss Function Comparison

Fig. 7 compares various bounding box regression losses on PCB defect detection. Standard GIoU [25] achieves a baseline of 96.7% mAP50. Individual advanced losses such as Wise-IoU [22] (96.5%), Focaler-IoU [23] (95.6%), and ShapeIoU [24] show varying performance when applied individually. Our WFS-IoU achieves the best performance at 97.3% mAP50 by explicitly modeling defect geometry through integration of all three components. The Wise component provides dynamic gradient allocation, Focaler emphasizes hard examples, and ShapeIoU accounts for aspect ratio variations in irregular PCB defects, outperforming single-component losses by 0.7–3.8 percentage points without increasing model complexity.

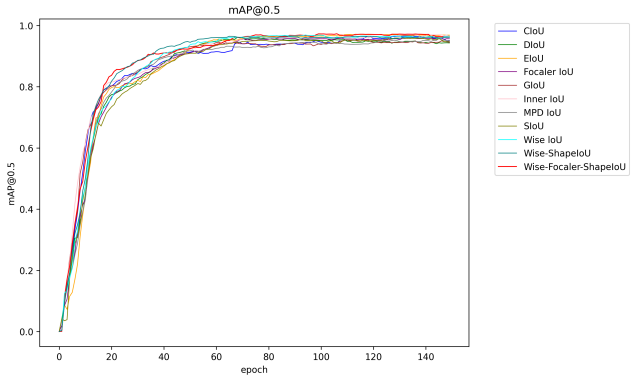


Fig. 7 Training curves for different loss functions.

3.4 Ablation Studies

Table 2 presents the contribution of each module to overall performance. Starting from the baseline RT-DETR-R18 that achieves 96.7% mAP50 with 19.9M parameters, we progressively integrate our proposed modules.

Module A (DSG Backbone) is the primary contributor to model compression, reducing parameters to 11.7M (41.2% reduction) and GFLOPs to 29.8 (47.7% reduction) while achieving 96.1% mAP50 with improved FPS of 112.5. This substantial reduction in model complexity is the foundation of our lightweight design. Module B (PME-Layer Encoder) achieves the highest individual accuracy at 97.3% mAP50, but with computational overhead (57.2 GFLOPs) and reduced inference speed (82.3 FPS) that limit practical deployment. Module C (DRBC3+wConv2d) further reduces computational cost through structural reparameterization while reaching 97.6% mAP50 with 110.2 FPS. Module D (WFS-IoU) enhances bounding box regression, boosting mAP50 to 97.1% with 109.0 FPS without any additional model parameters.

When combining modules, the lightweight DSG backbone introduces an accuracy-efficiency trade-off: A+B achieves 96.1% mAP50 with 40.2% fewer parameters (11.9M vs 19.9M) and 46.7% lower computational cost (30.4 vs 57.0 GFLOPs) at 115.4 FPS, but with reduced accuracy compared to Module B alone. The full model (A+B+C+D) recovers to 97.3% mAP50 through the contribution of WFS-IoU, while achieving 66.0% reduced computational cost (19.4 vs 57.0 GFLOPs), 48.7% fewer parameters (10.2M vs 19.9M), and real-time inference at 112.0 FPS. This demonstrates that the primary value of our design lies in achieving competitive accuracy with substantially improved efficiency, enabling deployment in resource-constrained scenarios where Module B alone would be infeasible.

3.5 Comparison with Other Detectors

Table 3 compares our method with mainstream detectors on the PKU-PCB dataset. Two-stage Faster R-CNN [1] achieves 93.6% mAP50 with 208.0 GFLOPs. The YOLO series shows varying performance: YOLOv5s [26] (96.9%), YOLOv8s [14] (96.2%), and YOLO11s [27] (95.9%). RT-DETR [12] variants demonstrate strong performance with R18 (96.7%), R34 (97.7%), R50 (98.1%), and R101 (98.7%).

DGFN-DETR achieves 97.3% mAP50 with 27.1% fewer parameters (10.2M vs 14.0M) and 57.5% lower

Table 2 Ablation study showing incremental improvements. A: DSG Backbone (replaces ResNet18), B: PME-Layer Encoder (replaces standard encoder), C: DRBC3+wConv2d (added modules), D: WFS-IoU Loss (replaces GIoU).

A	B	C	D	P (%)	R (%)	mAP50 (%)	mAP50-95 (%)	GFLOPs	Params (M)	FPS
				95.9	96.0	96.7	53.8	57.0	19.9	104.7
✓				95.8	96.1	96.1	52.8	29.8	11.7	112.5
	✓			97.2	97.7	97.3	55.4	57.2	20.0	82.3
		✓		97.4	97.0	97.6	55.9	46.3	18.2	110.2
			✓	96.1	96.1	97.1	54.0	57.0	19.9	109.0
✓	✓			96.3	95.3	96.1	53.3	30.4	11.9	115.4
✓	✓	✓		96.3	94.7	96.4	53.4	19.4	10.2	112.0
✓	✓	✓	✓	96.6	96.1	97.3	54.2	19.4	10.2	112.0

Table 3 Comparison with state-of-the-art methods on PCB defect dataset.

Method	P (%)	R (%)	mAP50 (%)	mAP50-95 (%)	GFLOPs	Params (M)
Faster R-CNN [1]	92.7	92.4	93.6	51.3	208.0	41.4
CenterNet [28]	90.8	90.7	91.8	50.0	22.0	14.2
YOLOv5s [26]	96.0	95.8	96.9	53.6	23.8	9.1
YOLOv8n [14]	96.0	92.8	95.7	52.2	8.1	3.0
YOLOv8s [14]	94.6	95.2	96.2	53.4	28.4	11.1
YOLO11n [27]	94.7	92.8	94.8	51.9	6.3	2.6
YOLO11s [27]	96.6	94.3	95.9	54.7	21.3	9.4
RT-DETR-R18 [12]	95.9	96.0	96.7	53.8	57.0	19.9
RT-DETR-R34 [12]	97.1	98.0	97.7	55.3	88.8	31.1
RT-DETR-R50 [12]	97.9	98.5	98.1	56.9	129.9	42.1
RT-DETR-R101 [12]	98.7	98.9	98.7	58.3	247.4	74.8
DGFN-DETR	96.6	96.1	97.3	54.2	19.4	10.2

computational cost (19.4 vs 45.6 GFLOPs) compared to recent Transformer-based approaches. Compared to RT-DETR-R34 which achieves similar accuracy (97.7%), our method uses 67.2% fewer parameters (10.2M vs 31.1M) and 78.2% lower GFLOPs (19.4 vs 88.8), making it far more suitable for edge deployment. Against the baseline RT-DETR-R18, DGFN-DETR achieves 0.6 percentage points higher mAP50 while reducing parameters by 48.7% and GFLOPs by 66.0%. Notably, among lightweight methods with fewer than 15M parameters, DGFN-DETR achieves the highest mAP50, demonstrating superior accuracy-efficiency trade-off. The results confirm that our lightweight design successfully compresses the model while improving detection accuracy, and the reduced computational cost naturally enables faster real-time inference.

3.6 Real-Time Performance Analysis

To comprehensively validate real-time deployment feasibility, Table 4 presents detailed performance profiling beyond basic throughput metrics, measuring latency distribution, memory footprint, and power consumption critical for industrial edge deployment.

DGFN-DETR achieves 8.9 ms average latency with 11.2 ms 95th percentile latency, providing deterministic timing guarantees essential for production line synchronization. The 39.2 MB model size (49.1% reduction) enables deployment on embedded platforms with limited resources. Power consumption of 68W (28.4% reduction) reduces thermal management requirements. These comprehensive metrics validate practical deployability on resource-constrained edge devices.

Table 4 Real-time performance analysis on NVIDIA RTX 4090.

Metric	RT-DETR-R18	DGFN-DETR
Average Latency (ms)	9.5	8.9
95th Percentile Latency (ms)	11.3	11.2
Model Size (MB)	77.0	39.2
Power Consumption (W)	95	68

3.7 Visualization Analysis

Figure 8 presents detection comparisons on challenging samples. DGFN-DETR accurately localizes small defects such as spurious copper and spurs that baseline methods miss or detect with lower confidence. For small defects, DGFN-DETR achieves higher confidence scores (0.81–0.83 vs 0.63–0.83 for RT-DETR) and improved detection rates, attributed to the WFS-IoU loss and enhanced multi-scale feature fusion. Confidence scores consistently exceed 0.82 across varying defect types, demonstrating robust detection capability.

Figure 9 shows activation heatmaps revealing each module's contribution. DGFN-DETR exhibits significantly more concentrated activation patterns around defect regions compared to RT-DETR's diffuse

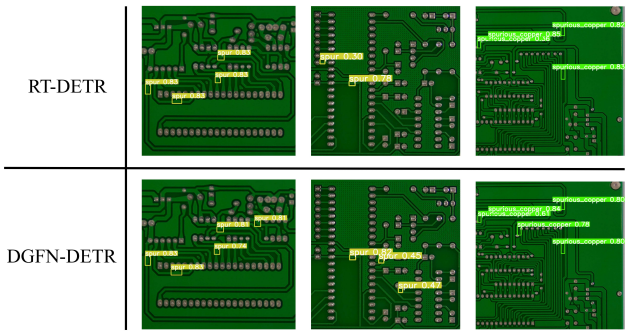


Fig. 8 Detection results comparison on PCB defect samples.

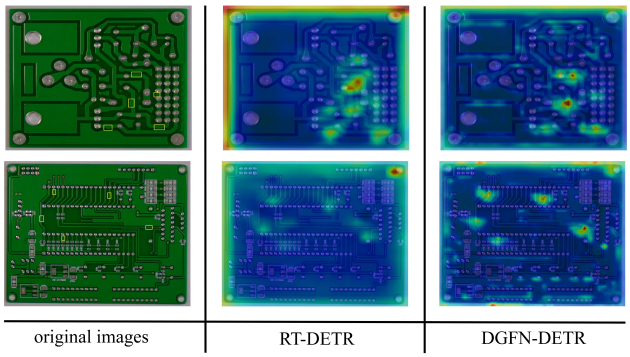


Fig. 9 Activation heatmap comparison between RT-DETR and DGFN-DETR.

attention distribution. The color gradient from blue (low activation) to red (high activation) clearly illustrates the model's focus. DGFN-DETR precisely focuses on spurious copper defects with intense red-yellow activation clusters while suppressing background regions. The Multi-Scale Adaptive Spatial Attention Gates dynamically weight dual pathway contributions, maintaining strong activation across defect locations with clearer boundaries between defect and non-defect regions.

3.8 Cross-Dataset Generalization

To validate the generalization capability of our lightweight architecture, we conducted cross-dataset evaluation on the DeepPCB dataset [37]. DeepPCB contains 1,500 image pairs with six defect types (open, short, mousebite, spur, copper, and pin-hole), providing a different data distribution from the PKU-PCB training set to assess cross-dataset performance.

As shown in Table 5, DGFN-DETR demonstrates superior generalization performance compared to the baseline RT-DETR-R18. On the DeepPCB dataset, our method achieves 97.85% mAP50, outperforming RT-DETR-R18 by 2.49 percentage points. More notably, DGFN-DETR shows substantial improvements in Precision (+2.63%) and Recall (+4.71%), indicating better detection robustness across different PCB manufacturing processes and imaging conditions.

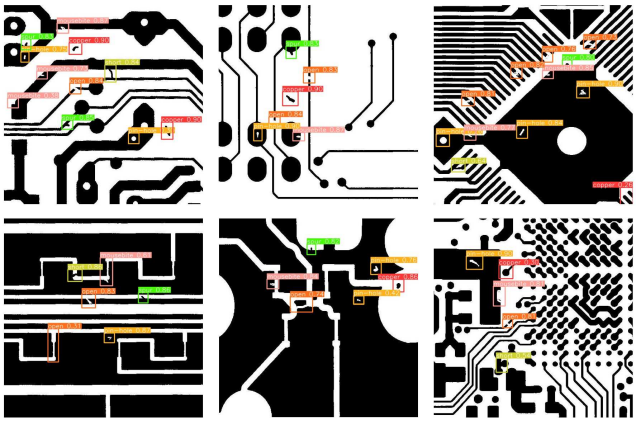


Fig. 10 Detection results on DeepPCB dataset demonstrating cross-dataset generalization capability.

Furthermore, our lightweight architecture maintains its efficiency advantage on the new dataset, achieving 127.9 FPS compared to 111.4 FPS for RT-DETR-R18 (14.8% improvement). This confirms that the computational benefits of our design are consistent across different deployment scenarios. The combination of improved accuracy and faster inference demonstrates that DGFN-DETR is well-suited for practical industrial deployment where models must generalize to diverse PCB types and manufacturing conditions.

4 Conclusion

This paper presents DGFN-DETR, a lightweight framework for industrial PCB defect detection that achieves significant model compression while maintaining high detection accuracy. The architecture integrates several key innovations: a Dual-Stream Gated Backbone for efficient feature extraction with substantially reduced parameters, a PME-Layer encoder with polarized attention achieving linear complexity for high-resolution feature processing, dilated reparameterized convolution blocks that collapse to single convolutions during inference thereby eliminating multi-branch overhead, weighted stride downsampling for information-preserving resolution reduction, and a geometry-aware WFS-IoU loss specifically designed for irregular defect shapes.

Experimental validation on the PKU-PCB dataset demonstrates that DGFN-DETR achieves competitive detection accuracy while significantly reducing both model parameters and computational cost compared to

Table 5 Cross-dataset generalization on DeepPCB dataset.

Method	P (%)	R (%)	mAP50 (%)	mAP75 (%)	FPS
RT-DETR-R18	94.79	89.51	95.36	61.83	111.4
DGFN-DETR	97.42	94.22	97.85	62.42	127.9

the baseline RT-DETR. The lightweight design naturally yields improved inference performance with reduced latency and deterministic timing guarantees, meeting real-time requirements for production line deployment. Comprehensive profiling reveals a compact 39.2 MB memory footprint and 68W power consumption, validating practical deployability on resource-constrained edge devices. Cross-dataset evaluation on DeepPCB further confirms the generalization capability of the proposed architecture across different PCB manufacturing processes and imaging conditions, demonstrating both improved accuracy and maintained efficiency advantages.

However, limitations remain: detection of extremely small defects (less than 2 pixels) presents challenges, and the model requires domain-specific training data for optimal performance on new PCB types. Future work will explore knowledge distillation for further model compression, quantization techniques for deployment on specialized edge accelerators, and semi-supervised learning to reduce annotation requirements. These directions aim to enhance the practical applicability of lightweight detection frameworks in diverse industrial scenarios while maintaining the balance between accuracy and efficiency established in this work.

Declarations

Funding

This work was supported by the Shanghai Artificial Intelligence Science Reform and Discipline Advancement Program (Grant No. 2024AI008).

Conflict of interest

The authors declare that they have no conflict of interest.

References

1. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems (NeurIPS), pp. 91–99 (2015)
2. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 779–788 (2016)
3. Ding, R., Dai, L., Li, G., Liu, H.: TDD-net: A tiny defect detection network for printed circuit boards. CAAI Trans. Intell. Technol. 4(2), 110–116 (2019)
4. Liu, Y., Zhang, X., Ding, Y., Guo, Q.: Automatic optical inspection for solder joint defects on PCB. IEEE Trans. Compon. Packag. Manuf. Technol. 10(3), 453–461 (2020)
5. Hu, B., Wang, J., Wang, J.: A novel method for PCB inspection based on image stitching. Meas. Sci. Technol. 31(7), 075401 (2020)
6. Tang, Y., Qiu, F., Jin, X., Zhang, X., He, Q.: Deep learning-based automatic defect identification of photovoltaic module using electroluminescence images. IEEE Trans. Ind. Inform. 19(2), 2018–2027 (2023)
7. Wang, H., Zhang, P., Zhao, Y.: Lightweight deep neural network for real-time PCB defect detection. Comput. Ind. Eng. **158**, 107430 (2021)
8. Chen, X., Lv, J., Fang, Y., Du, X.: Multi-scale feature fusion network for small defect detection. Measurement **187**, 110277 (2022)
9. Li, Q., Chen, W., Zhang, Y.: Attention-guided feature pyramid network for small object detection. Neurocomputing **493**, 14–24 (2022)
10. Yang, H., Zhang, L., Liu, X.: Adaptive feature aggregation framework for deep learning-based PCB defect classification. IEEE Trans. Instrum. Meas. **72**, 1–11 (2023)
11. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: Proc. Eur. Conf. Comput. Vis. (ECCV), pp. 213–229 (2020)
12. Lv, W., Xu, S., Zhao, Y., Wang, G., Wei, J., Cui, C., Du, Y., Dang, Q., Liu, Y.: DETRs beat YOLOs on real-time object detection. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 16965–16974 (2024)
13. Kim, S., Park, J., Lee, H.: Lightweight transformer for real-time industrial inspection. J. Manuf. Syst. **72**, 145–157 (2024)
14. Jocher, G., Chaurasia, A., Qiu, J.: YOLO by Ultralytics (2023). `https://github.com/ultralytics/ultralytics`
15. Kolahi, S.G., et al.: MSA^{2}Net: Multi-scale adaptive attention-guided network for medical image segmentation. In: Proc. British Machine Vis. Conf. (BMVC) (2024)
16. Meng, W., Luo, Y., Li, X., Jiang, D., Zhang, Z.: PolaFormer: Polarity-aware linear attention for vision transformers. In: Proc. Int. Conf. Learn. Represent. (ICLR) (2025)
17. Kong, L., Dong, J., Tang, J., Yang, M.-H., Pan, J.: Efficient visual state space model for image deblurring. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 12710–12719 (2025)
18. Yin, D., Hu, L., Li, B., Zhang, Y., Yang, X.: 5%>100%: Breaking performance shackles of full fine-tuning on visual recognition tasks. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 20071–20081 (2025)

19. Zhu, J., Chen, X., He, K., LeCun, Y., Liu, Z.: Transformers without normalization. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 14901–14911 (2025)
20. Ding, X., et al.: UniRepLKNet: A universal perception large-kernel ConvNet for audio, video, point cloud, time-series and image recognition. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR) (2024)
21. Cammarasana, S., Patanè, G.: Optimal weighted convolution for classification and denoising. arXiv:2405.24558 (2025)
22. Tong, Z., Chen, Y., Xu, Z., Yu, R.: Wise-IoU: Bounding box regression loss with dynamic focusing mechanism. arXiv:2301.10051 (2023)
23. Zhang, Y., et al.: Focaler-IoU: More focused intersection over union loss. arXiv:2401.10525 (2024)
24. Gevorgyan, Z.: SIoU loss: More powerful learning for bounding box regression. arXiv:2205.12740 (2022)
25. Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., Savarese, S.: Generalized intersection over union: A metric and a loss for bounding box regression. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 658–666 (2019)
26. Jocher, G.: YOLOv5 by Ultralytics (2020). `https://github.com/ultralytics/yolov5`
27. Jocher, G., Qiu, J.: YOLO11 by Ultralytics (2024). `https://github.com/ultralytics/ultralytics`
28. Duan, K., Bai, S., Xie, L., Qi, H., Huang, Q., Tian, Q.: CenterNet: Keypoint triplets for object detection. In: Proc. IEEE Int. Conf. Comput. Vis. (ICCV), pp. 6569–6578 (2019)
29. Chen, H., Wang, J., Chen, C., Shi, W., Huang, C., Xia, L., Xu, M., Xu, Z.: VanillaNet: The power of minimalism in deep learning. In: Advances in Neural Information Processing Systems (NeurIPS) (2023)
30. Woo, S., Debnath, S., Hu, R., Chen, X., Liu, Z., Kweon, I.S., Xie, S.: ConvNeXt V2: Co-designing and scaling ConvNets with masked autoencoders. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 16133–16142 (2023)
31. Cai, H., Li, J., Hu, M., Gan, C., Han, S.: EfficientViT: Lightweight multi-scale attention for on-device semantic segmentation. In: Proc. IEEE Int. Conf. Comput. Vis. (ICCV), pp. 17302–17313 (2023)
32. Chen, J., Kao, S., He, H., Zhuo, W., Wen, S., Lee, C.-H., Chan, S.-H.: Run, don't walk: Chasing higher FLOPS for faster neural networks. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 12021–12031 (2023)
33. Wang, C., Wu, Y., Deng, M., Li, K., Zhu, J., Shi, B., Wang, T., Wang, S., Sun, J., Lin, Y.: RepViT: Revisiting mobile CNN from ViT perspective. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 15909–15920 (2024)
34. Wen, X., Li, B., Wang, X., et al.: A Swin transformer-functionalized lightweight YOLOv5s for real-time coal–gangue detection. J. Real-Time Image Process. **20**, 47 (2023)
35. Li, Y., Hou, C.-Y., Li, J., Cheng, R., Cheng, H.-S., Chao, Y.-W., Chen, J.-J.: Large selective kernel network for remote sensing object detection. In: Proc. IEEE Int. Conf. Comput. Vis. (ICCV), pp. 16748–16759 (2023)
36. Li, H., Li, J., Wei, H., Liu, Z., Zhan, Z., Ren, Q.: Slim-neck by GSConv: a lightweight-design for real-time detector architectures. J. Real-Time Image Process. 21(3), 62 (2024)
37. Tang, S., He, F., Huang, X., Yang, J.: Online PCB defect detector on a new PCB defect dataset. arXiv:1902.06197 (2019)
38. Xiang, S., Fu, C., Song, W., Wang, X., Chen, J., Sham, C.-W.: FRT-DETR: Faster real-time end-to-end detector for industrial surface defects based on transformer. J. Real-Time Image Process. 22(5), 236 (2025)