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.

Light MV-DETR: Light MicroVision-DETR Based on RT-DETR for PCB Detection

Bingxiang Wang
Xi'an Jiaotong University
Mugen Zhou
Xi'an Jiaotong University
Wenzhuo Ma
Xi'an Jiaotong University
Tianyu Li
Lanzhou University of Technology
Changsheng Zhu
zhucs_2008@163.com

Lanzhou University of Technology

Research Article

Keywords:

Posted Date: November 25th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9012249/v1

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

Additional Declarations: No competing interests reported.

Light MV-DETR: Light MicroVision-DETR Based on RT-DETR for PCB Detection

Bingxiang Wang[1], Mugen Zhou[2†], Wenzhuo Ma[3†], Tianyu Li[4†], Changsheng Zhu[5*†]

[1]School of Computer Science and Technology, Xi'an Jiaotong University, Xi'an, Shaanxi, 710049, China.
[2]School of Software Engineering, Xi'an Jiaotong University, Xi'an, Shaanxi, 710049, China.
[3]School of Microelectronics, Xi'an Jiaotong University, Xi'an, Shaanxi, 710049, China.
[4]School of Computer Science and Artificial Intelligence, Lanzhou University of Technology, Lanzhou, Gansu, 730050,China.
[5*]School of Computer Science and Artificial Intelligence, Lanzhou University of Technology, Lanzhou, Gansu, 730050,China.

*Corresponding author(s). E-mail(s): zhucs_2008@163.com;
Contributing authors: 18042713825@163.com; mzhou2551@gmail.com;
1489127529@qq.com; Li_ty1482916370@163.com;
†These authors contributed equally to this work.

Abstract

High-resolution industrial vision plays a crucial role in PCB defect inspection, where dense small targets, weak texture contrast, and structural similarity pose significant challenges for accurate detection under real-time constraints. To address these issues, we propose Light-MicroVision-DETR, a lightweight end-to-end Transformer-based detection framework tailored for industrial defect inspection. The model integrates a MultiScale-LightNet backbone for efficient cross-scale feature representation, a Light-aware Fusion mechanism for enhanced multi-scale interaction, and a WaveletUnPool upsampling module to preserve fine structural details. Additionally, frequency-domain convolution expansion and structured spatial attention are introduced to improve robustness against noise and blurred boundaries. Experimental results on PCB defect datasets demonstrate that the proposed model achieves competitive accuracy with only 6.86M parameters and 13.7 GFLOPs, outperforming baseline RT-DETR while significantly reducing computational cost. The proposed framework achieves an effective balance between detection accuracy and efficiency, making it suitable for real-time industrial inspection scenarios.

Keywords: PCB defect detection, Multi-scale feature fusion, Frequency-domain convolution, Light-MicroVision-DETR

1 Introduction

Under the paradigm of Industry 4.0, high-resolution vision has become pivotal for quality control in electronics manufacturing. However, the miniaturization of Printed Circuit Boards (PCBs) introduces severe challenges for defect detection [1]. PCB images are characterized by dense geometric structures, where minute defects exhibit low contrast and high similarity to complex backgrounds [2]. This "weak texture, strong interference" environment often leads to semantic ambiguity between defects and circuits during feature extraction [3]. Furthermore, high-speed production lines impose stringent millisecond-level latency constraints. Consequently, detection systems face a critical trade-off: they must maintain high representation fidelity for micro-scale targets

while satisfying the rigorous real-time demands of edge deployment.

Despite substantial progress in object detection, mainstream frameworks face inherent bottlenecks when handling the minute, dense targets typical of PCBs. First, traditional feature pyramids suffer from information loss during downsampling, where the irreversible compression of high-frequency details blurs defect boundaries and hinders small-object localization. Second, CNN-based detectors are constrained by local inductive biases, limiting their ability to model the long-range dependencies required for complex circuit topologies. Conversely, while Transformers offer global modeling via self-attention, their quadratic computational complexity on high-resolution images prohibits efficient deployment on resource-constrained edge platforms. Furthermore, naive feature stacking often induces computational redundancy, failing to achieve discriminative robustness within a limited parameter budget.

Addressing these challenges, this paper proposes a lightweight end-to-end Transformer model tailored for minute and dense defect detection, namely Light-MicroVision-DETR. Furthermore, we have constructed a closed-loop detection system facilitating synergy from the industrial site to the cloud (Fig. 1). First, high-performance event-based cameras are utilized to capture high-dynamic-range real-time images on the production line, after which the data stream is asynchronously transmitted to an edge-cloud server for end-to-end inference. In the cloud, Light-MV-DETR achieves second-level determination of minute defects. Finally, detection results are fed back to the factory terminal in real time to assist technicians in performing precise interactive defect correction. This deployment mode ensures the system's operability and closed-loop quality control within actual industrial environments.

More importantly, considering multi-dimensional complex interferences in actual factory environments, this paper not only validates the model's detection efficacy on standard assembly lines but also conducts targeted robustness assessments under extreme conditions. Experimental data indicate that even under extreme operating conditions such as poor lighting, camera defocus blur, low-light, out-of-focus scenarios, and composite interferences, Light-MicroVision-DETR maintains superior detection stability. While maintaining extremely low computational costs and parameter counts, this solution significantly enhances the system's discriminative accuracy under non-ideal imaging conditions, providing highly reliable and deployable technical support for minute object detection technology in complex and variable real-world smart manufacturing scenarios.

The main contributions of this paper can be summarized as follows:

(1) A lightweight backbone, MultiScale-LightNet, is designed to achieve efficient semantic modeling through AKConv-based cross-scale enhancement and EMA multi-scale attention. This significantly improves the quality of multi-scale features with an extremely low computational load, providing a robust feature foundation for real-time detection.

(2) A cross-scale feature fusion pyramid composed of WaveletUnPool, Light-aware Fusion, and WTConv2d is designed, which significantly enhances the lightweight network's capability in edge restoration for minute and dense defects as well as in multi-scale feature modeling.

(3) A front-end platform integrating real-time monitoring and human–machine collaboration is established. Through stability verification under extreme environments, the practical deployment value of the system under complex factory operating conditions is demonstrated.

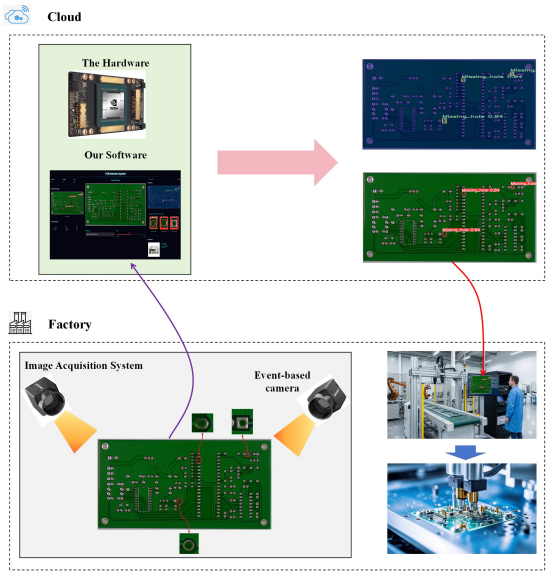

Fig. 1 Industrial deployment diagram of Light-MV-DETR.

2 Related Works

The evolutionary trajectory of object detection technology clearly delineates the transition from traditional machine learning to deep representation learning, exhibiting a recent trend toward a paradigm shift from Convolutional Neural Networks (CNNs) to Transformer architectures. During the era dominated by CNNs, algorithmic evolution was primarily categorized

into high-precision two-stage detection, represented by Faster R-CNN [4], and high-efficiency one-stage detection, represented by the YOLO series. Through continuous architectural evolution, the YOLO series has progressed from YOLOv5 [5], which solidified performance via multi-scale feature pyramids and lightweight designs, to YOLOv7 [6], which improved engineering efficiency through programmable gradients and task-decoupled structures, and subsequently to YOLOv8 [7], which introduced an anchor-free architecture to enhance model universality. Recently, YOLOv9 [8] strengthened object distribution modeling via an NMS-free end-to-end training strategy, while YOLOv10 [9] further optimized inference latency through systematic structural re-parameterization.However, as noted in numerous studies on PCB defects, traditional CNNs, constrained by local receptive fields, often suffer from insufficient feature response when handling defect scenarios characterized by complex topological variations and long-range dependencies.

With the emergence of the Transformer [10] architecture, object detection has been reshaped as an end-to-end set prediction problem. The DETR model explicitly models cross-region dependencies via global self-attention mechanisms, thoroughly eliminating reliance on handcrafted components such as anchors and NMS. Addressing the slow convergence and high computational costs of the vanilla DETR, subsequent research such as Deformable DETR [11] introduced sparse deformable attention mechanisms, while Anchor-DETR [12] leveraged anchor priors to accelerate model alignment. DN-DETR [13] significantly promoted training stability by introducing a denoising training strategy. The advent of RT-DETR [14], which combines efficient attention mechanisms with hybrid feature processing paths, enabled Transformer detectors to achieve real-time inference capabilities comparable to the YOLO series for the first time while retaining global modeling advantages. Although these methods perform exceptionally well in general object detection, a key focus in current academic research remains how to reduce feature correlation costs in high-resolution images while maintaining the global perception advantages of Transformers, particularly when facing extremely small and dense detection tasks such as those in PCBs.

To address the challenges of dense small-object detection, researchers have undertaken extensive targeted modifications to both YOLO-series and DETR-series models. Some works focus on enhancing multi-scale feature consistency or introducing edge-detail enhancement modules to alleviate information loss during downsampling, while others improve localization accuracy in dense

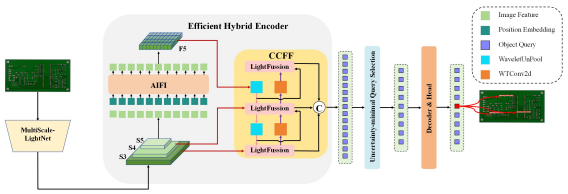

Fig. 2 The architecture of Light-MV-DETR. The proposed network integrates Multi-Scale LightNet and an efficient hybrid encoder with CCF fusion to enhance feature representation while maintaining lightweight design.

scenes by constructing spatial priors or decoupled attention mechanisms. However, existing works often overlook the potential of frequency-domain information in image detail restoration, leading to an inevitable loss of high-frequency image features in the pursuit of lightweight designs.Building upon the integration of the local inductive bias of CNNs and the global modeling advantages of Transformers, this work further explores the deep coupling of frequency-domain transforms and Transformer architectures. By constructing a frequency-aware lightweight detection framework, this study not only fills the gap in fine-grained feature inversion within existing models but also provides a new theoretical basis for achieving high-performance industrial detection under extreme computational constraints.

3 METHODOLOGY

As illustrated in Fig. 2, Light-MicroVision-DETR advances the RT-DETR baseline through five key structural innovations. First, the MultiScale-LightNet backbone employs primary–auxiliary dual-branch convolutions to synergize sparse spatial and dense channel features, resolving the semantic weakness of depthwise separable convolutions while maintaining efficient feature extraction. Second, an Efficient Hybrid Encoder alternates between lightweight self-attention and multi-scale convolutions, reconciling global dependency modeling with local geometric sensitivity. To enhance feature fusion, the CCFF module performs cascaded cross-scale aggregation via Distance-Aware Fusion, suppressing background noise while maintaining high-frequency fidelity. Furthermore, WaveletUnPool is adopted to recover directional details via inverse Haar transforms, ensuring sub-pixel structural alignment. Finally, WTConv2d fuses wavelet-domain frequency reconstruction with spatial residuals, optimizing edge separability with negligible computational overhead.

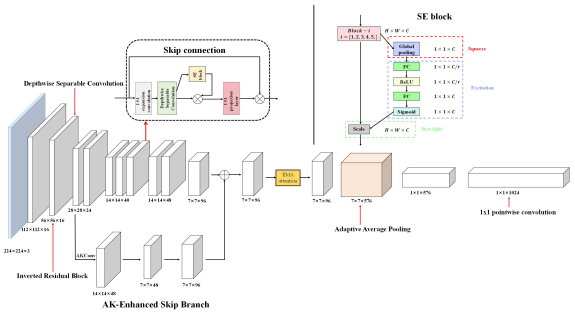

Fig. 3 The architecture of MultiScale-Light Net.

3.1 MultiScale-Light Net

To reconcile the trade-off between feature expressiveness and computational efficiency in real-time detection, we propose MultiScale-LightNet, a lightweight backbone architecture. As shown in Fig. 3, by incorporating the AK-Enhanced Skip Branch and Enhanced Multi-scale Attention (EMA), the network enhances cross-scale modeling capabilities with minimal parameter overhead. The architecture begins with lightweight spatial compression, followed by pointwise convolution for cross-channel reorganization. Given input features $X \in \mathbb{R}^{H\times W\times C}$, the overall mapping is formulated as follows:

$$F_1 = \mathrm{ReLU}\big(W_{1\times1}\big(W^{\mathrm{dw}}_{3\times3}(I)\big)\big) \tag{1}$$

In the initial stages, inverted residual bottlenecks are stacked to progressively expand channel dimensions and receptive fields. Given input feature F, the module performs high-dimensional expansion via 1×1 convolution, extracts local structural details via depthwise convolution, and restores dimensions through projection. This design optimizes the multiply–accumulate (MAC) budget by efficiently applying nonlinearity, thereby enhancing channel discriminability. To augment multi-scale context, the architecture integrates an AK-Enhanced Skip Branch. Central to this branch is AKConv, which leverages a learnable kernel alignment mechanism to amplify responses from broader spatial regions. Its mathematical formulation is defined as:

$$X_{\mathrm{out}}(p) = \sum_{k=1}^{n} G_k(p)\cdot X(p+\Delta p_k) \tag{2}$$

Subsequently, channel alignment is performed via a 1×1 convolution, followed by element-wise addition. This parameter-efficient fusion enriches deep semantic features with cross-scale structural priors, thereby enhancing robustness for dense small targets. The fused features then undergo

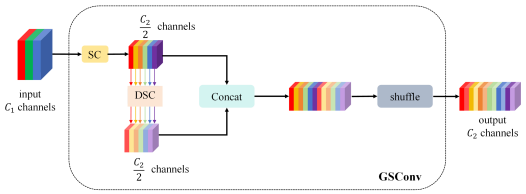

Fig. 4 The architecture of GSConv.

resolution-sensitive refinement via the EMA module. By utilizing three parallel depthwise convolutions with varying receptive fields, EMA aggregates multi-scale spatial cues to generate the response:

$$M = \sum_{s\in\{1,2,3\}} \alpha_s \cdot \mathrm{DWConv}_{k_s}(F_{\mathrm{fuse}}) \tag{3}$$

Here, α denotes learnable weights for adaptive scale regulation. Subsequently, EMA performs channel-wise recalibration by applying global pooling followed by a bottleneck MLP to generate the weight vector w. This mechanism significantly enhances semantic discriminability in deep, low-resolution layers with negligible latency overhead.

To bolster high-level semantics, Squeeze-and-Excitation (SE) modules are integrated into intermediate and late-stage blocks. By explicitly modeling channel dependencies, SE suppresses irrelevant features while amplifying informative ones. Unlike direct high-dimensional nonlinearity, SE leverages global statistics for robust weight estimation, effectively mitigating semantic degradation in lightweight networks.

Finally, MultiScale-LightNet yields multi-scale features $\{S_3, S_4, S_5\}$, which are projected to a fixed dimension d_{model} via 1×1 convolution for Transformer compatibility. By synergizing inverted residuals, AKConv, EMA, and SE, the backbone achieves a superior balance between computational efficiency and structural fidelity, ensuring stable detection performance for dense small targets.

3.2 GSConv

To enhance convolutional efficacy under lightweight constraints, we incorporate the GSConv [15] module. As depicted in Fig. 4, this structure reconciles computational efficiency with feature integrity via a synergistic dual-branch design. The primary branch employs Standard Convolution (SC) to generate $\frac{c_2}{2}$ channel-dense, high-rank features, ensuring semantic stability. Conversely, the auxiliary branch utilizes Depthwise Separable Convolution (DSC) to capture spatial textures for the remaining channels.A subsequent channel shuffle facilitates uniform feature mixing. Distinct from ShuffleNet, this operation

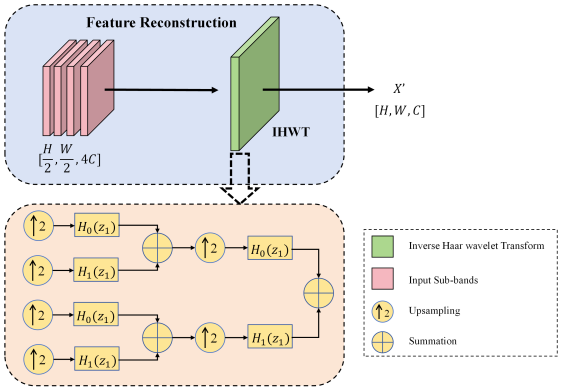

Fig. 5 The architecture of WaveletUnPool.

functions as semantic injection: dense features from the primary branch compensate for the inherent channel sparsity of the auxiliary branch, yielding a balanced, non-fragmented hybrid distribution. Consequently, GSConv achieves structural bistability—preserving semantic consistency via the primary path while expanding spatial reception through the auxiliary path—formulated as follows:

$$X_{\text{out}} = \text{Shuffle}(X_{\text{SC}} \oplus X_{\text{DSC}}) \tag{4}$$

The adaptability of GSConv varies distinctly across network stages. In shallow layers, characterized by high spatial resolution and low semantic density, GSConv mitigates the semantic degradation often associated with extensive DSC usage. Conversely, in deep layers where spatial compression and channel redundancy peak, it preserves local geometric and global semantic consistency while curbing the exponential computational costs of Standard Convolution. Beyond low FLOPs, GSConv's core advantage lies in its structural synergy: the SC branch compensates for the reduced channel correlation inherent in DSC, enhancing texture representation, while the shuffle operation ensures robust cross-stream information flow. Crucially, when integrated with Transformer encoders, GSConv provides essential local structural priors. This bridges the gap between convolutional and attentional domains, optimizing feature propagation and sharpening decision boundaries. Ultimately, GSConv serves as a robust unit that reconciles structural integrity with lightweight constraints, offering an optimal trade-off for real-time detection.

3.3 WaveletUnPool

The fidelity of feature recovery during upsampling is critical in real-time object detection. Conventional methods, such as bilinear interpolation and deconvolution, rely on empirical assumptions lacking reversible structural constraints. Consequently, they function as approximate inferences, often inducing artifacts such as blurring, ghosting, and boundary drift, while failing to preserve spatial topology or compensate for directional textures.To address these limitations, we propose WaveletUnPool, inspired by WaveletPool [16]. As can be observed in Fig. 5, by shifting the paradigm from empirical interpolation to reversible reconstruction, WaveletUnPool utilizes the inverse Haar wavelet transform. This introduces a strict orthogonal reconstruction mechanism, endowing the process with clear mathematical interpretability. Leveraging the compact support, completeness, and orthogonality of Haar wavelets, the relevant functions are defined as follows:

$$\phi_1(x) = \frac{1}{\sqrt{2}}\,\phi_{1,0}(x) + \frac{1}{\sqrt{2}}\,\phi_{1,1}(x) \tag{5}$$

$$\psi_1(x) = \frac{1}{\sqrt{2}}\,\phi_{1,0}(x) - \frac{1}{\sqrt{2}}\,\phi_{1,1}(x) \tag{6}$$

$$\phi_0(x) = \begin{cases} 1, & 0 \le x < 1, \\ 0, & \text{otherwise.} \end{cases} \tag{7}$$

In the two-dimensional domain, discrete wavelet decomposition partitions features into four subbands: the approximation component A (low-frequency) and three high-frequency components (H, V, D) capturing horizontal, vertical, and diagonal variations. This transformation trades spatial resolution for a fourfold channel expansion while preserving strict reversibility.Leveraging this property, WaveletUnPool reformulates upsampling as an inverse wavelet reconstruction task, effectively mitigating the blurring and artifacts inherent to unconstrained interpolation. The module utilizes these sub-bands to independently recover global structure and directional textures. By employing inverse filters $H_0(z^{-1})$ and $H_1(z^{-1})$ derived from the Haar orthogonal basis, it executes zero-insertion, convolution, and cross-subband summation to generate a resolution-enhanced output, defined as:

$$X_{\text{rec}} = \text{IWT}(A, H, V, D) \tag{8}$$

Here, IWT denotes the two-dimensional inverse wavelet transform. By explicitly preserving and integrating the three directional high-frequency sub-bands (H, V, D), this structure imparts superior edge sensitivity and local structural discriminability to the upsampled features. Consequently, WaveletUnPool effectively addresses the challenge of detecting PCB defects, which are typically characterized by slender geometries and low contrast. Unlike conventional interpolation that often leads to feature smoothing, this approach preserves the integrity of minute defects, ensuring robust detection.

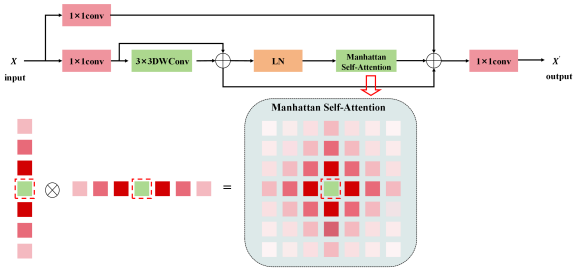

Fig. 6 The architecture of Distance-Aware Fusion.

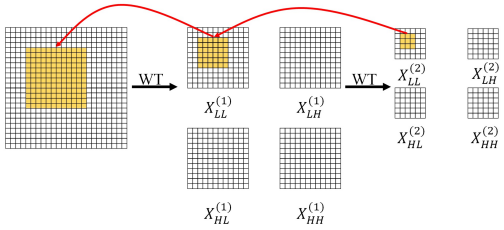

Fig. 7 The architecture of Distance-Aware WTConv2d.

3.4 Light Fusion

To unify convolutional local geometric priors with explicit spatial decay attention, we draw inspiration from RetBlock [17] and propose the Light-aware Fusion structure. As presented in Fig. 6, While convolutions excel at extracting local textures and self-attention captures long-range dependencies, the lack of coupled geometric constraints often hinders structural consistency in dense small-object scenarios. Addressing this issue, we introduce a Manhattan distance-based spatial decay prior to regulate the attention scope, synergized with a convolutional branch for stable local bases. This design enables cross-scale modeling with approximately linear complexity. The module comprises three synergistic functional paths. First, input features $X \in \mathbb{R}^{H\times W\times C}$ are projected to a compact subspace via 1×1 convolutions, establishing a unified semantic basis while reducing computational load. Second, a primary residual path utilizes cross-layer skip connections to ensure stable gradient flow and preserve high-level semantics. Concurrently, a convolution enhancement path captures shallow neighborhood interactions, providing reliable local geometric priors to ground subsequent long-range modeling.

Finally, in the attention path, the Manhattan distance explicitly constrains spatial interactions via a decay matrix. For coordinate pairs (x_n, y_n) and (x_m, y_m), the decay weight is defined as:

$$D_{nm}^{2d} = \gamma^{|x_n - x_m| + |y_n - y_m|} \tag{9}$$

$$\mathrm{MaSA}(X) = \left(\mathrm{Softmax}(QK^T) \odot D^{2d}\right) V \tag{10}$$

This formulation induces an exponential decay in feature similarity relative to spatial distance, effectively attenuating long-range noise while prioritizing local structural integrity. To efficiently compute two-dimensional decay attention, we employ an axial decomposition strategy, decoupling the process into orthogonal horizontal and vertical components. The resulting one-dimensional attention mechanisms and final output are defined as:

$$\mathrm{Attn}_H = \left(\mathrm{Softmax}\left(Q_H K_H^T\right) \odot D^H\right) \tag{11}$$

$$\mathrm{Attn}_W = \left(\mathrm{Softmax}\left(Q_W K_W^T\right) \odot D^W\right) \tag{12}$$

$$\mathrm{MaSA}(X) = \mathrm{Attn}_H \left(\mathrm{Attn}_W V\right)^T \tag{13}$$

The three branches are uniformly aggregated via a 1×1 convolution, synthesizing local geometry, long-range dependencies, and spatial priors into a unified representation. Channel shuffling is subsequently applied to equilibrate the feature distribution across subspaces, thereby optimizing cross-branch interaction efficiency. Meanwhile, the residual path guarantees semantic stability during deep propagation. Crucially, the explicit spatial decay mechanism within Distance-Aware Fusion confers high sensitivity to geometric shapes, boundary contours, and directional textures. This attribute significantly enhances discriminability compared with traditional fusion methods, particularly in scenarios characterized by dense targets and substantial scale variations.

3.5 WTConv2d

To circumvent the inherent limitations of standard convolutions—where parameter counts grow quadratically with kernel size—and to simultaneously achieve low latency and high representational capability in real-time scenarios, we introduce WTConv2d [18]. As depicted in Fig. 7, this module shifts the convolution operation from the spatial domain to the multi-scale wavelet domain. Leveraging the intrinsic advantages of wavelet decomposition in multi-scale modeling, this design significantly expands the effective receptive field while maintaining linear parameter scaling, effectively endowing lightweight networks with modeling capacity typically reserved for large-kernel convolutions.

Theoretically grounded in the two-dimensional orthogonal wavelet transform, WTConv2d employs fixed depthwise filters to decompose

input features into four sub-bands: a low-frequency structural component and three directional high-frequency components. With a stride of 2, the decomposition causes spatial resolution to decrease exponentially at each recursive level, facilitating progressive aggregation of global structural information within low-frequency bands. Unlike direct large-kernel applications in high-resolution space, this wavelet-based approach achieves broader coverage at lower computational cost, providing a structured frequency-domain basis for cross-scale representation.To ensure representational consistency, local convolutions are applied exclusively within individual sub-bands, followed by the inverse wavelet transform for linear fusion. This process ensures strict alignment between structural enhancement, frequency-domain interaction, and spatial reconstruction.

The continuous multi-scale enhancement mechanism, comprising recursive wavelet decomposition, local convolutional processing, and layer-wise inverse transformation, is formulated as follows:

$$[X_{LL}, X_{LH}, X_{HL}, X_{HH}] = \mathrm{WT}(X) \tag{14}$$

$$Y = \mathrm{IWT}(\mathrm{Conv}(W, \mathrm{WT}(X))) \tag{15}$$

$$[X_{LL}^{(i)}, X_{LH}^{(i)}, X_{HL}^{(i)}, X_{HH}^{(i)}] = \mathrm{WT}\left(X_{LL}^{(i-1)}\right) \tag{16}$$

$$Z^{(i)} = \mathrm{IWT}\left(Y_{LL}^{(i)}, Y_{LH}^{(i)}, Y_{HL}^{(i)}, Y_{HH}^{(i)}\right) + Z^{(i+1)} \tag{17}$$

$$Z^{(0)} = Y_{LL}^{(0)} + Z^{(1)} \tag{18}$$

This fusion paradigm enables deep low-frequency sub-bands to incorporate extensive spatial context at minimal computational cost, leveraging progressive resolution reduction. Simultaneously, high-frequency sub-bands preserve directional textures and local variations, allowing WTConv2d to achieve a natural equilibrium between structural sensitivity and global dependency modeling. Ultimately, the recursive inverse wavelet reconstruction linearly aggregates convolutional responses across varying scales and frequencies into a unified spatial coordinate system. Consequently, WTConv2d demonstrates superior structural expressiveness, global receptive capability, and parameter efficiency compared with traditional convolutions, providing a mathematically rigorous and practically viable solution for constructing efficient multi-scale vision backbones.

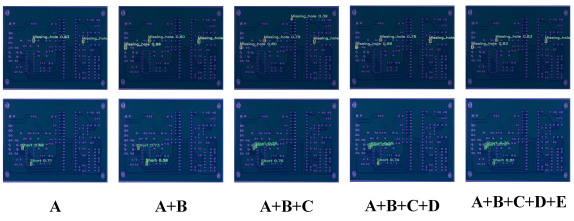

Fig. 8 Ablation Experiment Diagram.

4 EXPERIMENTS

4.1 Dataset

The PCB Defects dataset serves as a prominent benchmark in the field of industrial vision, specifically tailored for evaluating fine-grained recognition in high-resolution small-object scenarios. The dataset encompasses six distinct defect categories: open circuits, short circuits, mouse bites, missing holes, spurs, and missing copper. These anomalies present significant detection challenges due to their minute spatial scale, substantial morphological variability, and high textural similarity to complex circuit backgrounds.Furthermore, the dataset is characterized by low contrast and imaging noise, often causing traditional models to suffer from missed detections or false positives. Consequently, PCB Defects establishes a rigorous standard for verifying the robustness and feature representation capabilities of lightweight, high-precision detection frameworks. To ensure experimental standardization and reproducibility, the dataset is partitioned into training, validation, and test subsets with a ratio of 7:2:1, providing a stable basis for assessing model generalization and accuracy.

4.2 Experimental Setup

All experiments in this paper were conducted on an NVIDIA A100 GPU. The operating environment consisted of the Ubuntu operating system, a 20-core Intel(R) Xeon(R) Platinum 8470Q processor, 50 GB of RAM, and 40 GB of GPU memory. The models were implemented in Python 3.12 using PyTorch 2.3.0 with CUDA 12.1. The input image resolution was set to 640×640.The AdamW optimizer was employed for model training. The training configuration was set to 200 epochs with a batch size of 32, an initial learning rate of 1×10^{-4}, and a weight decay of 1×10^{-4}.

4.3 Ablation Studies

We performed incremental ablation experiments to verify the efficacy of modules A through E. The backbone MultiScale-LightNet (A), while efficient, demonstrated limited capacity for feature representation in dense, minute defect scenarios.

Table 1 Ablation Study

A	B	C	D	E	Params/M	GFLOPs	P	R	FPS	$AP_{IoU=0.5}$	$AP_{IoU=0.5:0.95}$
					19.87	57.0	89.2	89.1	59	89.9	43.8
1					9.61	25.1	78.7	72.4	76	90.7	44.6
1	1				8.45	21.5	87.1	75.1	79	91.9	45.3
1	1	1			8.35	20.7	89.6	88.6	80	93.1	46.4
1	1	1	1		8.22	20.5	92.6	91.5	81	95.7	51.2
1	1	1	1	1	6.86	13.7	93.1	94.0	87	95.8	51.9

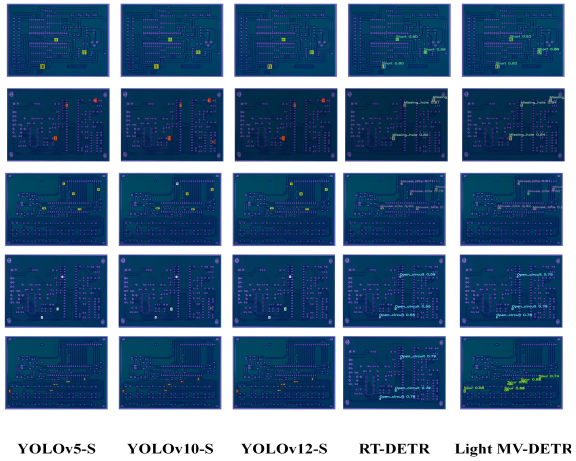

Fig. 9 Comparison Experiment Heatmap.

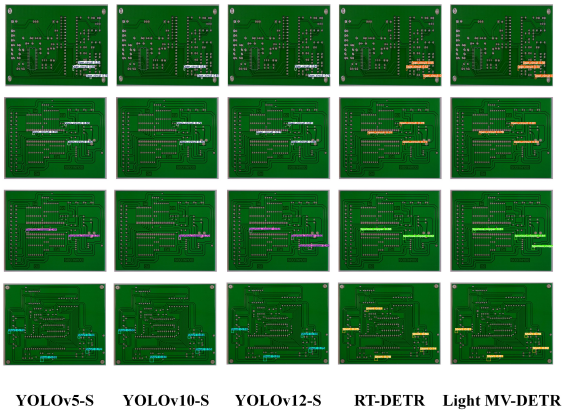

Fig. 10 Comparison Experiment Detection Image.

The introduction of GSConv (B) addressed this by enhancing channel coupling, compensating for the semantic sparsity of depthwise convolutions and improving local structural fidelity. Building on this, WaveletUnPool (C) was added to rectify upsampling artifacts; utilizing the inverse wavelet transform, it reconstructs high-frequency edge details, thereby refining boundary sensitivity.

The capability for global modeling was significantly bolstered by WTConv2d (D), which expands the receptive field through frequency-domain decomposition, enabling the capture of cross-scale geometric consistency amidst complex backgrounds. The architecture was finalized with Light-Aware Fusion (E), which harmonizes global and local features. This module leverages Manhattan decay attention to attenuate remote noise while strengthening local geometric continuity via dual-path aggregation. Collectively, these components construct a robust feature optimization pipeline, where complementary improvements in semantic correlation, structural reconstruction, and spatial attention synergize to maximize detection performance under strict computational budgets. As demonstrated in Table 1 and Fig. 9, the progressive integration of these modules consistently enhances detection accuracy.

4.4 Comparative Experiments

Comparative evaluations on the PCB defect detection benchmark elucidate distinct performance disparities between Light MicroVision-DETR and mainstream frameworks, notably the YOLO series. While YOLO architectures excel in inference speed and feature pyramid efficiency, their standard convolutional backbones reveal inherent limitations in sensitivity toward fine-grained defects. This deficiency is exacerbated by high-similarity texture interference, frequently resulting in feature smoothing and boundary blurring.

Conversely, Light MicroVision-DETR overcomes these bottlenecks by synergistically integrating wavelet frequency-domain convolution, directional high-frequency reconstruction, and structured spatial attention. This architecture endows the model with robust global dependency modeling and precise texture analysis capabilities with minimal computational overhead. Consequently, it outperforms competing detectors across precision, recall, and stability metrics, as shown in Table 2, Fig. 9, and Fig. 10. Notably, despite a significantly reduced parameter count and lower GFLOPs, our model maintains—and often surpasses—the performance of heavier counterparts. These findings underscore an effective unification of low complexity and high accuracy, validating the method's structural superiority and practical viability for industrial deployment.

Table 2 Comparison with different object detection methods.

Method	Params/M	GFLOPs	P	R	FPS	$AP_{IoU=0.5}$	$AP_{IoU=0.5:0.95}$
Faster R-CNN	41.14	63.26	85.8	82.1	47	85.8	43.6
RetinaNet [19]	20	63.7	86.1	83.1	40	86.1	43.5
FCOS [20]	32.15	68.52	87.0	83.2	42	87.0	44.3
SSD [21]	23.61	68.28	86.3	82.6	46	86.3	43.2
Sparse R-CNN [22]	106.2	129.5	87.2	81.9	45	87.2	44.7
Cascade R-CNN [23]	69.1	205.3	87.2	81.9	45	87.0	44.5
YOLOv5-S	9.11	23.8	95.5	89.7	59	92.3	48.8
YOLOv8-S	11.13	28.4	93.8	88.3	62	92.7	49.4
YOLOv10-S	7.22	21.4	93.4	84.3	61	92.5	47.2
YOLOv11-S [24]	9.41	21.3	97.2	87.8	56	94.8	47.6
YOLOv12-S [25]	9.07	18.7	95.3	88.7	58	94.6	47.6
YOLOv13-S [26]	9.00	20.0	93.8	85.6	61	90.9	44.9
RT-DETR	19.87	57.0	89.2	89.1	59	89.9	43.8
RT-DETRv2-M [27]	20	60	91.8	89.0	78	91.8	42.6
DN-DETR	48	195	90.5	86.5	54	90.5	41.9
DINO-DETR [28]	47	279	91.2	88.0	61	91.2	43.0
Light-MV-DETR	**6.86**	**13.7**	93.1	**94.0**	**87**	**95.8**	**51.9**

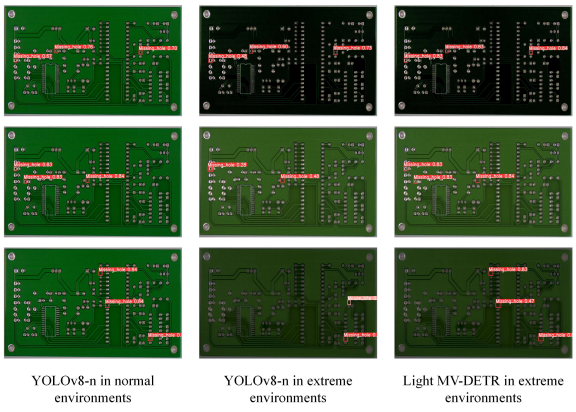

Fig. 11 Extreme Environment Test Diagram.

4.5 Test In Extreme Environments

To verify real-world reliability, we employed diffusion models to synthesize physically plausible extreme environments, including low-illumination, dynamic defocus blurring, and compound interference. These synthetic datasets challenge detectors with non-linear noise and high-frequency information loss, simulating worst-case industrial scenarios.

In comparative benchmarks against YOLOv8-n, distinct performance gaps were observed, as shown in Fig. 11. As shown in the extreme environment test diagram, YOLOv8-n exhibits a high rate of missed detections when identifying minute defects with blurred edges, primarily due to the limitations of standard convolutions in processing degraded high-frequency features. Conversely, Light-MicroVision-DETR demonstrated superior robustness, achieving zero missed detections across all simulated conditions. This performance is underpinned by the synergistic effect of the WaveletUnpool module and LightFusion: the former ensures the recovery of fine-grained edge details via wavelet reconstruction, while the latter preserves global semantic consistency. This confirms the model's capability to overcome 'feature annihilation,' providing a solid foundation for deployment in volatile industrial settings.

Table 3 FPS test under low-performance device

Equipment	Model	FPS
i5-12400F + RTX3060Ti	YOLOv5	33
	YOLOv8	32
	YOLOv10	33
	YOLOv11	30
	YOLOv12	35
	RT-DETR	37
	Light-MV-DETR	**60**

4.6 Efficiency and Real-Time Performance

We further evaluated the model's robustness in resource-constrained settings through a stress test focused on inference stability. Given that industrial edge gateways often operate under significant computational loads, minimizing latency is a prerequisite for deployment. By throttling GPU power and bandwidth to simulate typical edge constraints, we analyzed the system's temporal consistency when handling high-resolution inputs.

Attributed to the lightweight design of the MultiScale-LightNet backbone and the efficient feature processing of GSConv, Light MV-DETR demonstrates exceptional resilience in low-performance environments. The model sustains a stable 60 FPS output with minimal variance, effectively mitigating the computational bottlenecks

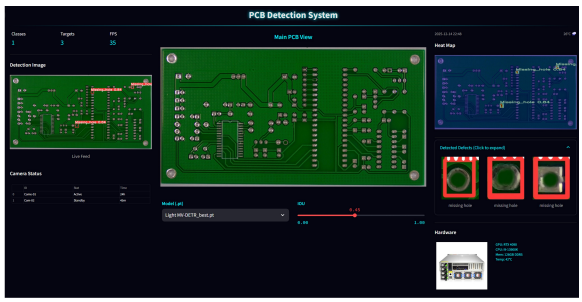

Fig. 12 Factory deployment software.

common in embedded deployment, as evidenced in Table 3. This performance ensures that the system maintains real-time fidelity even during stochastic data spikes or hardware instability. Crucially, this low-latency characteristic supports the seamless integration of cloud inference with interactive correction interfaces, validating the model's suitability for high-throughput, collaborative industrial ecosystems.

4.7 Industrial Deployment

To substantiate the portability and engineering utility of Light MV-DETR in industrial production, we developed and deployed an integrated monitoring prototype based on a cloud-edge synergistic architecture. As illustrated in Fig. 12, the system employs a responsive Web-based front-end designed to provide operators with high-transparency, real-time feedback and interactive quality control mechanisms. The core logic centers on the real-time fidelity of PCB defect detection.

The interface features a top-layer status monitoring module that dynamically synchronizes cloud inference results, displaying key metrics such as total defect counts, average confidence scores, and high-frame-rate (FPS) streams. To address the perceptibility challenges associated with minute and dense targets, the system utilizes coordinate data to render defect heatmaps and precise annotations on high-definition source images. This visual guidance significantly mitigates the difficulty of identifying sub-pixel anomalies, such as missing holes or short circuits. Furthermore, a built-in multi-dimensional statistical module characterizes defect distribution patterns via time-series histograms, offering objective data support for process parameter optimization. Crucially, the system implements an efficient correction mechanism. Through the bottom interaction panel, operators can retrieve magnified, enhanced views of specific regions to rapidly verify cloud-based determinations. This solution validates the proposed model's unique capability to reconcile high-performance recognition with low-resource deployment in complex operational environments.

5 Conclusion

In response to the persistent bottlenecks in high-resolution industrial vision—characterized by dense small objects, low texture contrast, and high structural similarity—this study establishes Light MicroVision-DETR as a systematic solution for PCB defect detection. By re-engineering the feature hierarchy of RT-DETR from a high-level modeling perspective, the proposed framework constructs a synergistic mechanism that integrates local geometric enhancement, fine-grained frequency-domain reconstruction, and cross-scale global modeling. This design effectively compensates for the structural deficiencies inherent in traditional lightweight frameworks, specifically regarding high-frequency detail fidelity and global semantic consistency.

Empirical evaluations demonstrate that the representational complementarity arising from multi-module synergy not only significantly enhances sensitivity to fine-grained defects but also validates the structural stability of frequency-domain convolution and explicit spatial attention in industrial scenarios. Quantitatively, the model surpasses multiple mainstream lightweight detectors in accuracy while achieving a substantial reduction in parameter count and computational overhead. Qualitatively, visualizations reveal superior boundary reconstruction and noise suppression capabilities. These findings underscore the critical role of high-frequency directional recovery and structured spatial constraints in bolstering semantic focus and geometric consistency under weak-texture conditions. Notably, the ability of WTConv2d to exponentially expand the effective receptive field within a lightweight profile offers a scalable frequency-domain modeling pathway for hybrid convolution-attention architectures.

Looking ahead, future research will focus on three strategic directions: (1) Adaptive Modeling: Introducing dynamic convolutions, learnable wavelet bases, or structural re-parameterization to enable frequency-domain modules to adaptively accommodate diverse defect morphologies; (2) Edge Efficiency: Integrating low-bit quantization and structured sparsification to maximize the energy efficiency ratio of wavelet convolutions and explicit attention on edge devices; and (3) Cross-Domain Extension: Extending this framework to cross-modal or temporal tasks to verify its geometric consistency in broader industrial contexts. Ultimately, the structured modeling philosophy embodied by Light MicroVision-DETR provides a generalizable system paradigm for dense small object detection, laying a robust methodological foundation for the future of lightweight intelligent perception in industry.

6 Declarations

6.1 Funding

This paper was supported by Cooperation project between Lanzhou University of Technology and Longnan Power Supply Company of State Grid Gansu Electric Power Company.

References

[1] W. Jiang, T. Li, S. Zhang, W. Chen, and J. Yang, "PCB defects target detection combining multi-scale and attention mechanism," *Eng. Appl. Artif. Intell.*, vol. 123, p. 106359, Aug. 2023, doi: 10.1016/j.engappai.2023.106359.

[2] S. Cong, Z. Shang, and Q. Huang, "Detection for printed circuit boards (PCBs) delamination defects using optical/thermal fusion imaging technique," *Infrared Phys. Technol.*, vol. 127, p. 104399, Dec. 2022, doi: 10.1016/j.infrared.2022.104399.

[3] D. Makwana, S. C. T. R., and S. Mittal, "PCBSegClassNet—A lightweight network for segmentation and classification of PCB component," *Expert Syst. Appl.*, vol. 225, p. 120029, Sep. 2023, doi: 10.1016/j.eswa.2023.120029.

[4] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2015.

[5] "YOLOv5: A deep look into the internal features of the popular object detector," arXiv:2407.20892, 2025.

[6] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," in *Proc. CVPR*, 2023, pp. 7464–7475.

[7] M. Yaseen, "What is YOLOv8: An in-depth exploration of the next generation object detector," arXiv:2408.15857, 2024.

[8] C.-Y. Wang, I.-H. Yeh, and H.-Y. M. Liao, "YOLOv9: Learning what you want to learn using programmable gradient information," in *Computer Vision – ECCV 2024*, LNCS, vol. 15089. Springer, 2025, pp. 1–21.

[9] A. Wang et al., "YOLOv10: Real-time end-to-end object detection," 2024.

[10] A. Vaswani et al., "Attention is all you need," arXiv:1706.03762, 2017.

[11] X. Zhu et al., "Deformable DETR: Deformable transformers for end-to-end object detection," arXiv:2010.04159, 2021.

[12] Y. Wang et al., "Anchor DETR: Query design for transformer-based object detection," arXiv:2109.07107, 2022.

[13] F. Li et al., "DN-DETR: Accelerate DETR training by introducing query denoising," arXiv:2203.01305, 2022.

[14] Y. Zhao et al., "DETRs beat YOLOs on real-time object detection," arXiv:2304.08069, 2024.

[15] H. Li et al., "Slim-neck by GSConv: A lightweight design for real-time detector architectures," *Real-Time Image Process.*, vol. 21, no. 3, p. 62, Jun. 2024.

[16] G. Xu et al., "Haar wavelet downsampling: A simple but effective downsampling module for semantic segmentation," *Pattern Recognit.*, vol. 143, p. 109819, Nov. 2023.

[17] Q. Fan et al., "RMT: Retentive networks meet vision transformers," arXiv:2309.11523, 2025.

[18] S.-E. Finder et al., "Wavelet convolutions for large receptive fields," arXiv:2407.05848, 2024.

[19] T.-Y. Lin et al., "Focal loss for dense object detection," arXiv:1708.02002, 2018.

[20] Z. Tian et al., "FCOS: Fully convolutional one-stage object detection," arXiv:1904.01355, 2019.

[21] W. Liu et al., "SSD: Single shot multibox detector," in *ECCV*, 2016.

[22] P. Sun et al., "Sparse R-CNN: End-to-end object detection with learnable proposals," arXiv:2011.12450, 2020.

[23] Z. Cai and N. Vasconcelos, "Cascade R-CNN: Delving into high quality object detection," arXiv:1712.00726, 2017.

[24] "YOLOv11: An overview of the key architectural enhancements," arXiv:2410.17725, 2025.

[25] "YOLOv12: Attention-centric real-time object detectors," arXiv:2502.12524, 2025.

[26] M. Lei et al., "YOLOv13: Real-time object detection with hypergraph-enhanced adaptive visual perception," arXiv:2506.17733, 2025.

[27] W. Lv et al., "RT-DETRv2: Improved baseline with bag-of-freebies for real-time detection transformer," arXiv:2407.17140, 2024.

[28] H. Zhang et al., "DINO-DETR with improved denoising anchor boxes for end-to-end object detection," arXiv:2203.03605, 2022.