

Article

MulPViT-SimAM: An Electronic Substrate Defect Detection Framework for Addressing Class Imbalance Problems

Yuting Wang ¹, Liming Sun ² , Bang An ¹ and Ruiyun Yu ^{1,*}¹ Software College, Northeastern University, Shenyang 110819, China; wangyt@stumail.neu.edu.cn (Y.W.)² School of Computer Science and Engineering, Northeastern University, Shenyang 110169, China

* Correspondence: yury@mail.neu.edu.cn

Abstract

As the cornerstone of contemporary electronics, the quality of electronic substrates—including Printed Circuit Boards (PCBs) and Ceramic Package Substrates (CPSs)—is intrinsic to product reliability. However, automated inspection is currently impeded by two persistent obstacles: the drastic multi-scale variation in defects and the acute class imbalance within defect datasets. Conventional deep learning approaches often fail to reconcile these challenges simultaneously, leading to suboptimal recognition of rare defect categories. To bridge this gap, we propose Multi-scale Partial Vision Transformer—Simple, Parameter-free Attention Module (MulPViT-SimAM), a robust framework designed for class-imbalanced electronic substrate defect detection. Our method features a novel multi-scale backbone (MulPViT) that synergizes partial convolutions with hierarchical attention mechanisms, facilitating the efficient extraction of both fine-grained local textures and global contextual dependencies. Additionally, we embed the Simple, Parameter-free Attention Module (SimAM) into the feature fusion stage to adaptively highlight defect-specific features while dampening background noise. To further mitigate data imbalance, we utilize the Equalized Focal Loss (EFL) function, which employs a category-specific modulating factor to dynamically equilibrate the learning focus across different classes. Comprehensive benchmarking reveals state-of-the-art performance, achieving mAP@0.5 scores of 95.7% on the standard PKU-MARKET-PCB dataset and 54.2% on the highly challenging CPS2D-AD dataset. Significantly, our approach effectively mitigates class imbalance, narrowing the performance deviation of rare categories to just 4.3% on the PKU-Market-PCB dataset and 1.4% on the CPS2D-AD dataset, compared to 11.8% and 7.5% in baseline models. These findings position MulPViT-SimAM as a viable and efficient solution for industrial quality control.

Keywords: electronic substrate defect detection; vision transformer; class imbalance; partial convolution; feature fusion; SimAM attention



Academic Editor: Ahmed Abu-Siada

Received: 9 March 2026

Revised: 14 April 2026

Accepted: 16 April 2026

Published: 20 April 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

As the cornerstone of contemporary electronic devices, electronic substrates such as Printed Circuit Boards (PCBs) and Ceramic Package Substrates (CPSs) provide essential electrical connectivity and mechanical stability for components. With the swift advancement of electronics manufacturing, quality standards for electronic substrates have escalated significantly. Surface anomalies—ranging from missing holes, mouse bites, and open circuits to shorts, spurious copper, and spurs—pose serious risks of product malfunction and potential safety hazards [1]. Conventional reliance on manual visual inspection proves

inadequate for high-throughput production settings due to its labor-intensive nature, inefficiency, and susceptibility to human error [2]. Consequently, the pursuit of automated, intelligent approaches for electronic substrate defect identification has emerged as a pivotal objective in industrial quality assurance.

In recent years, deep learning-based object detection frameworks have demonstrated substantial efficacy and are extensively utilized in electronic substrate inspection tasks [3]. Notably, the You Only Look Once (YOLO) family of single-stage detectors has garnered widespread preference for delivering an optimal trade-off between detection precision and inference latency [4]. Contemporary research has explored various enhancements to the YOLO architecture, including the integration of attention mechanisms, optimization of feature pyramid networks, and the deployment of lightweight backbones [5,6]. For instance, Xu et al. proposed ESMNet, an enhanced YOLOv7-based framework incorporating multi-scale feature fusion and attention mechanisms for surface defect detection in precision metal workpieces, demonstrating the viability of transformer-augmented YOLO architectures in real industrial settings [7]. Similarly, Shen et al. introduced a multi-expert diffusion model that addresses class imbalance through generative data augmentation for surface defect detection in special control valve equipment, underscoring the importance of tackling rare-category recognition in safety-critical manufacturing [8]. Nevertheless, processing the distinctive characteristics of electronic substrate defect imagery remains a formidable hurdle for these methodologies; compounding this challenge, existing research efforts in this field have been largely confined to PCBs, with far fewer automated detection studies dedicated to CPS.

A primary obstacle involves the significant multi-scale heterogeneity of electronic substrate defects. These anomalies vary drastically in morphology and dimension, spanning from minute spurs occupying merely a few pixels to extensive short circuits covering hundreds [9]. Standard Convolutional Neural Networks (CNNs) are constrained by fixed receptive fields, which hinders their capacity to simultaneously extract features across diverse scales. While Feature Pyramid Networks (FPNs) attempt to mitigate this by handling multi-scale detection, the semantic disparity between pyramid levels often results in the loss of critical information during feature fusion [10].

For the multi-scale feature extraction challenge addressed earlier, Vision Transformers (ViTs) have emerged as a promising alternative to CNNs in visual recognition, distinguished by their self-attention mechanisms that excel at capturing long-range dependencies [11]. Contemporary research has successfully adapted ViTs for object detection, yielding impressive results on standard benchmarks [12]. Notably, NVIDIA's FasterViT architecture synergizes the global modeling strengths of ViTs with the efficient local feature extraction of CNNs via a hierarchical attention strategy, delivering state-of-the-art accuracy with optimized computational overhead [13]. Nevertheless, directly deploying ViT-based methods in industrial defect inspection faces two practical barriers: their substantial computational overhead and their dependence on large-scale labeled training datasets.

A secondary but equally critical obstacle is the issue of class imbalance within electronic substrate defect datasets. In actual manufacturing environments, defect categories exhibit highly disproportionate frequencies, where certain types are ubiquitous while others are scarce [14]. This uneven distribution biases detection models towards majority classes, frequently causing them to overlook rare yet crucial defects. While Focal Loss was introduced to mitigate foreground-background imbalance in dense detection by suppressing easy samples [15], it employs a uniform modulating factor across all classes, failing to account for the heterogeneous imbalance levels inherent to different defect types.

To counteract class imbalance, diverse strategies involving re-weighting and re-sampling have been explored [16]. Although class-balanced loss functions attempt to

calibrate weights based on sample counts, they lack the flexibility to adapt to dynamic training progressions [17]. Conversely, Equalized Focal Loss (EFL) [18], originally proposed for dense object detection, utilizes a category-specific modulating factor to dynamically tune the focusing degree based on the imbalance severity of each class. While EFL is an established methodological approach, its potential has not been fully realized in industrial quality control. In this work, we introduce an application-based innovation by adapting EFL to the context of electronic substrate defect detection. This specific industrial domain features unique and severe long-tailed distributions that easily confound standard detectors. By empirically validating EFL within our framework, we demonstrate a robust applied solution to these dynamic imbalance patterns.

Attention mechanisms are extensively utilized in object detection architectures to augment feature representation capabilities [19]. Distinct from conventional approaches, the Simple, Parameter-free Attention Module (SimAM) derives 3D attention weights via a neuroscience-inspired energy function, notably devoid of any additional trainable parameters [20]. Empirical evidence across diverse defect detection domains highlights SimAM's proficiency in bolstering feature discrimination while preserving computational economy [21]. Consequently, integrating SimAM into feature fusion networks holds substantial promise for identifying minute defects amidst intricate backgrounds.

To mitigate computational redundancy within neural networks, Partial Convolution (PConv) was introduced, a technique that restricts convolution operations to a specific subset of input channels [22]. FasterNet, which is founded on PConv architecture, delivers superior FLOPS (floating-point operations per second) relative to depthwise convolution-based models, rendering it highly viable for real-time industrial deployment. Nevertheless, achieving an optimal equilibrium between computational efficiency and feature representation remains a critical imperative for defect detection tasks, particularly where the capture of fine-grained details is paramount.

To transcend the inherent limitations of pure YOLO-based and ViT-based architectures, the rationale behind our framework design lies in a synergistic combinatorial advance tailored for substrate inspection. Existing YOLO detectors, primarily CNN-based, offer rapid inference but struggle with the global contextual awareness required for varying defect scales. Conversely, pure ViT models capture global dependencies but impose prohibitive computational burdens unsuitable for real-time industrial deployment. Our architecture bridges this gap by combining the complementary strengths of these paradigms. The proposed MulPViT backbone selectively utilizes Partial Convolutions (PConv) for ultra-efficient local texture extraction at shallow layers, reserving ViT's hierarchical attention for deep global context modeling. Furthermore, to counteract the potential background noise amplified by global attention, we embed the parameter-free SimAM module in the fusion stage. This specific combination creates a mutually reinforcing mechanism: MulPViT secures comprehensive multi-scale representations at low computational costs, while SimAM adaptively suppresses complex background interference, thereby achieving an optimal equilibrium between detection precision and real-time efficiency that surpasses conventional singular-paradigm models.

In this study, we present MulPViT-SimAM, a holistic framework designed to mitigate the dual challenges of multi-scale feature extraction and class imbalance in electronic substrate defect detection. The primary contributions of this work are summarized as follows:

- (1) We engineered MulPViT, a multi-scale feature extraction backbone that synergizes Partial Convolution blocks in the initial layers for rapid local feature processing with hierarchical attention modules in the deeper layers for global context modeling. This hybrid architecture strikes an optimal trade-off between computational economy and feature representational power.

- (2) We embedded the SimAM attention mechanism within the feature fusion network. This module adaptively highlights defect-specific features while attenuating background noise, thereby boosting detection precision for minute and low-contrast anomalies.
- (3) We utilized the Equalized Focal Loss (EFL) function, augmented by a gradient-guided dynamic adjustment strategy. This approach effectively counteracts class imbalance, significantly enhancing the recognition of rare defect categories without compromising the accuracy of frequent ones.
- (4) Comprehensive empirical evaluations were performed on the PKU-Market-PCB and CPS2D-AD benchmarks, the two mainstream substrate types constituting electronic packaging devices. The results confirm that our method achieves state-of-the-art performance in both accuracy and inference speed, establishing its viability for automated industrial quality inspection of electronic substrates.

2. Methodology

2.1. Overview of Network Architecture

Building upon the YOLOX architecture [23], the proposed algorithm is a refined single-stage object detection framework tailored for electronic substrate defect identification. Unlike conventional models that rely solely on CNNs or ViTs, our system derives its strength from a synergistic combination. It employs an optimized multi-scale Vision Transformer (ViT) network (MulPViT) to efficiently balance local and global feature extraction, while the feature fusion neck is augmented by the integration of the SimAM module to filter out substrate background noise without increasing parameter complexity. Target localization and categorization are executed via a decoupled head structure, which separates classification and regression tasks; the complete architectural schematic is depicted in Figure 1. Functionally, the input image is first processed by the enhanced ViT Backbone to extract multi-scale features. Subsequently, these features are synthesized by the fusion module across the P3, P4, and P5 scales. The resulting fused output feature maps are P3_OUT, P4_OUT, and P5_OUT. These are then passed through the decoupled head prediction structure to obtain the target position regression (Reg), confidence output (IoU), and classification output (Cls).

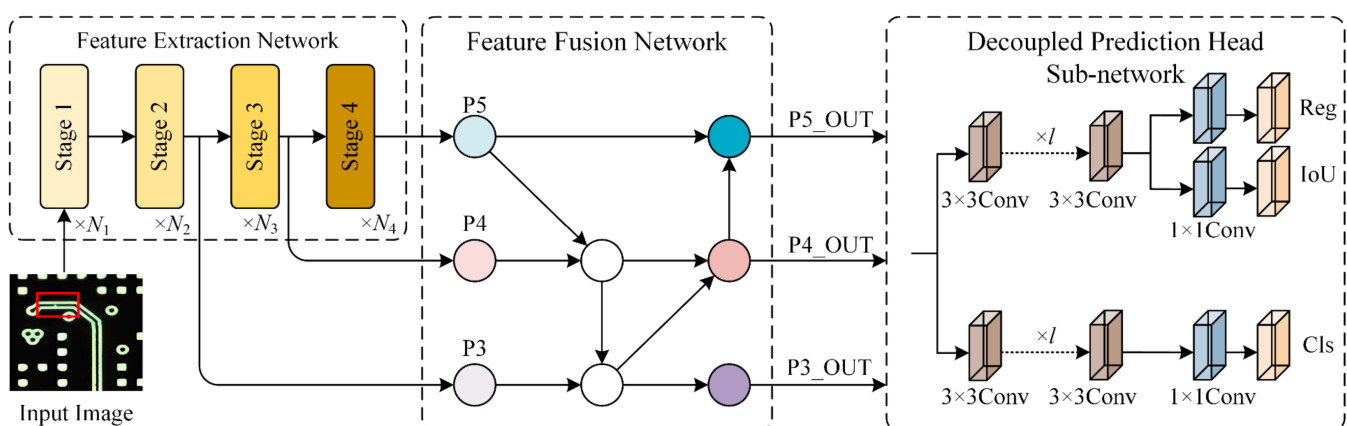


Figure 1. MulPViT-SimAM network architecture.

2.2. Multi-Scale Feature Extraction Backbone Network

Developed by NVIDIA, FasterViT serves as a hardware-centric neural backbone that synergizes the global context modeling of ViTs with the local feature extraction strengths of CNNs [24]. While the standard FasterViT relies on conventional convolutional blocks, this study introduces a generalized multi-scale feature extraction network, termed MulPViT, by

integrating Partial Convolution (PConv) [25] to mitigate computational complexity during defect detection. The architecture of MulPViT is depicted in Figure 2.

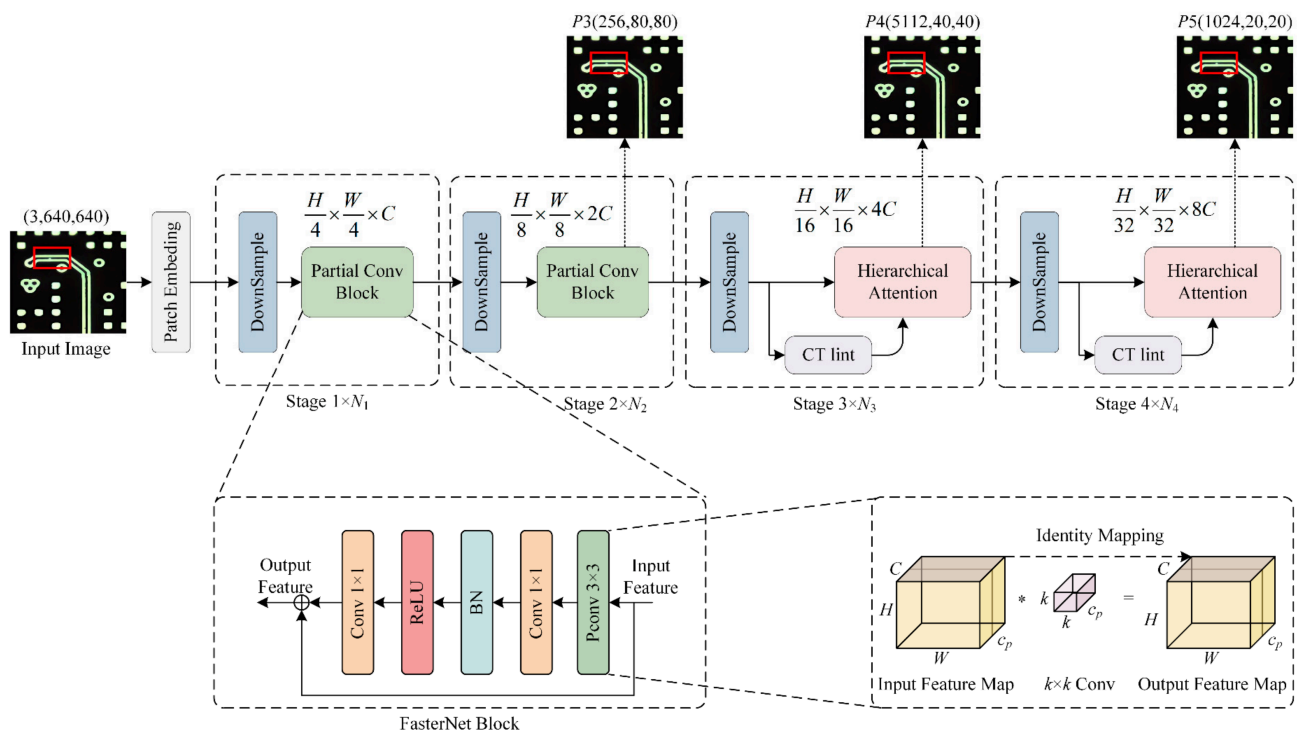


Figure 2. MulPViT multi-scale feature extraction network.

By fusing partial convolutional layers with ViT attention mechanisms, MulPViT aims to streamline feature extraction and reduce the model complexity, thereby reducing parameter count and enhancing computational efficiency. To accommodate target objects of varying dimensions, MulPViT employs a four-stage process for multi-scale feature extraction. Specifically, the feature maps generated from Stages 2, 3, and 4 constitute the primary multi-scale outputs of the backbone.

In the initial Stages 1 and 2, Partial Conv Blocks are the dominant extraction units. As referenced in [26], each block comprises two 1×1 convolutional layers, a ReLU activation, a Batch Normalization layer, and a PConv layer, linked via residual connections. To address the channel redundancy inherent in CNN feature maps, PConv is adopted to curb computational costs. Its operational mechanism, illustrated in the bottom right of Figure 2, involves performing standard convolution on only a subset of input channels while applying identity mapping to the rest. This strategy optimizes both memory access and computational load without compromising essential feature information. Assuming the height and width of the input feature map are H and W respectively, with c channels, c_p consecutive channels are taken as representative of the input feature map. Feature extraction is performed on these first c_p channels using a convolution with a stride of 1 and padding of $(\text{kernel_size} - 1)/2$. The computational complexity (FLOPs) of PConv is expressed as in Equation (1):

$$FLOPs = H \times W \times k^2 \times c_p^2 \quad (1)$$

The partial ratio is defined as $r = c_p/c$; when $r < 1$, the FLOPs of PConv are only r^2 times those of conventional convolution. In terms of memory access, PConv reduces the

volume to r times that of conventional convolution, with Memory_PConv estimated by Equation (2):

$$\text{Memory_PConv} = H \times W \times 2c_p + k^2 \times c_p^2 \approx H \times W \times 2c_p \quad (2)$$

In Stages 3 and 4, feature extraction follows the design of FasterViT, utilizing a ViT structure based on hierarchical attention. The working principle, illustrated in Figure 3, introduces carrier tokens (CTs) to aggregate and transmit global information. Interactions between local and global information are established by connecting local tokens from local windows with the global CTs. Because a local window only accesses its corresponding CTs, the computational overhead for information exchange is reduced, improving the efficiency of the ViT network in industrial environments. By alternating between global and local window token strategies, a hierarchical attention network is constructed, allowing the model to focus on both global contextual information and local details simultaneously.

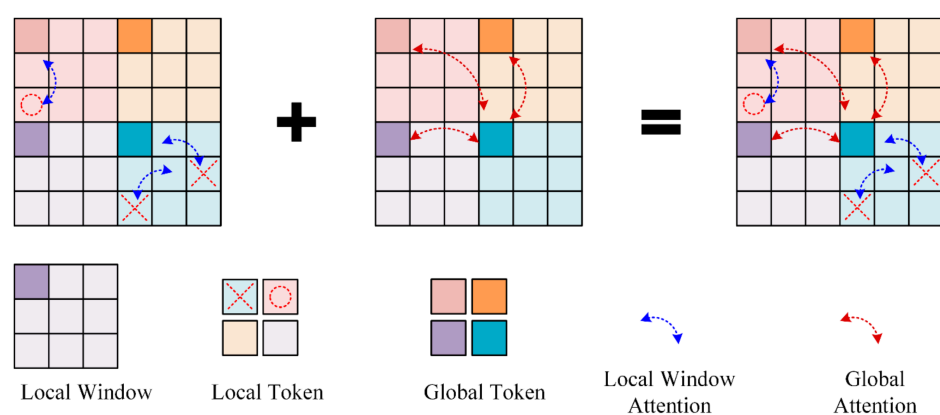


Figure 3. Information interaction pattern of the hierarchical attention mechanism.

Before detailing the mathematical formulations, the core mechanism of our hierarchical attention can be intuitively understood as establishing an efficient ‘communication bridge’ between local micro-textures and global macro-structures. In traditional ViTs, calculating attention across all pixels simultaneously incurs massive computational overhead. To streamline this, we introduce global Carrier Tokens (CTs). Intuitively, one can think of local window tokens as ‘inspectors’ focused on specific, tiny patches of the electronic substrate, while CTs act as ‘supervisors’ that summarize the overall layout of the board. Instead of every inspector talking to every other inspector, they only exchange information with the supervisors (CTs). This localized-to-global interaction drastically reduces computational complexity while ensuring that fine-grained defect features are analyzed within the correct global context.

To enable local attention with manageable complexity, the feature map is partitioned into non-overlapping windows of fixed size $k \times k$. The total number of windows $n \times n$ is determined by the spatial dimensions H, W and the window size k :

$$x_1^d = \text{split}_{n \times n}(x), n = \frac{H \times W}{k^2} \quad (3)$$

In the equation, H and W represent the height and width of the input feature map respectively, d is the number of channels, and k is the window size. A 3×3 convolution is used to encode the feature map to construct global tokens, as shown in Equation (4):

$$x_c^e = \text{Conv}_{3 \times 3}(x) \quad (4)$$

Each local window's features are then compressed into a single Carrier Token (CT) via average pooling. This CT serves as a compact summary of the window's spatial content, enabling efficient global information exchange without attending to every individual token:

$$x_{ct}^e = AvgPool_{H \times W \rightarrow n^2 L}(x_c^e) \quad (5)$$

The CTs are first updated among themselves using multi-head self-attention (MHSA) to aggregate global information across all windows. This is analogous to a 'global meeting' where each window representative (CT) updates its understanding based on all others. A subsequent MLP with GeLU activation enriches the representation with non-linear transformations:

$$x_{ct}^{e1} = x_{ct}^e + \gamma_1 \odot MHSA(LN(x_{ct}^e)) \quad (6)$$

Apply a multilayer perceptron (MLP) and GeLU activation function for post-processing to enable the network to better capture contextual information; the calculation process is shown in Equation (7):

$$x_{ct}^{e2} = x_{ct}^{e1} + \gamma_2 \odot MLP(LN(x_{ct}^{e1})) \quad (7)$$

In the equations, x_{ct}^e represents the original un-updated CTs; MHSA denotes the multi-head self-attention operation; and x_{ct}^{e1} represents the CTs weighted through multi-head self-attention. MLP refers to the operation of introducing non-linear fitting through the GeLU activation function and an MLP, which is used to further process the updated CTs and enhance the representation capability of the network. x_{ct}^{e2} denotes the weighted CTs after the non-linear fitting via the MLP. γ_1 and γ_2 are learnable weight parameters. LN stands for the layer normalization operation, used to normalize attention weights.

Next, the modeling capability of the network for long-range and short-range spatial feature information is enhanced through the interaction between the local window tokens and the global CTs. First, the local window tokens and the global CTs are concatenated. Then, the concatenated tokens are updated through the multi-head self-attention operation, followed by further weighted updates through the MLP operation to obtain the final tokens. The calculation process of this interaction is shown in Equations (8)–(10):

$$x_w^e = Concat(x_1^0, x_{ct}^{e2}) \quad (8)$$

$$x_w^{e1} = x_w^e + \gamma_3 \odot MHSA(LN(x_w^e)) \quad (9)$$

$$x_w^{e2} = x_w^{e1} + \gamma_4 \odot MLP(LN(x_w^{e1})) \quad (10)$$

In the equations, x_w^e is the token obtained through concatenation; x_w^{e1} is the token after weighting by multi-head self-attention; and x_w^{e2} is the token further weighted after MLP processing. γ_3 and γ_4 are also learnable weight parameters.

Through the aforementioned operations, the local window tokens and global carrier tokens are allowed to influence each other, which helps to better capture important features and contextual information in the image. The updated tokens are then split into local window tokens and global carrier tokens. x_1^0 represents the local window tokens, which consist of the first L tokens of the token vector. $x_{ct,1}^e$ represents the global carrier tokens, which consist of the subsequent tokens starting from the L -th token of the token vector. After the joint attention step, the merged token sequence is split back into local window tokens and updated CTs. Local tokens are reshaped into spatial feature maps, while CTs

are upsampled and added back via residual connection, infusing global context into the locally processed features:

$$(x_1^\delta, x_{ct,1}^\varepsilon) = \text{split}(x_w^{\varepsilon 2}) \quad (11)$$

The final hierarchical attention output is expressed as shown in Equation (12):

$$x = \text{upsample}_{n^2 L \rightarrow H \times W}(x_{ct,1}^\varepsilon) + \text{merge}_{n^2 k^2 \rightarrow H \times W}(x_1^\delta) \quad (12)$$

To ensure that the network retains the positional information of features in the spatial dimension, absolute positional bias is introduced, and a dual-layer MLP is employed to embed the absolute positional bias information into the features. Hierarchical attention divides self-attention into two components: global self-attention and local self-attention. Meanwhile, through interaction with the carrier tokens, global and local self-attention enable the network to better integrate global and local information, thereby further enhancing the network's feature extraction capability.

2.3. Attention-Enhanced Feature Fusion

The human visual system possesses an innate ability to isolate salient details within an image while filtering out irrelevant background noise. To replicate this selective focus and mitigate the confusion arising from the high visual similarity between defects and the substrate background, we integrated the Simple, Parameter-free Attention Module (SimAM) [27] into the neck fusion network. SimAM computes a 3D attention weight for each neuron based on the principle of spatial suppression from neuroscience: neurons that strongly inhibit their spatial neighbors are considered more salient and assigned higher importance weights. Formally, the importance of a target neuron t is measured via the following energy function, which quantifies how distinctly t differs from its surrounding neurons on the same channel:

$$e_t(w_t, b_t, y, x_i) = \frac{1}{M-1} \sum_{i=1, i \neq t}^M (-1 - (w_t x_i + b_t))^2 + (1 - (w_t t + b_t))^2 + \lambda w_t^2 \quad (13)$$

In the equation, M represents the number of neurons on a channel; λ is a weight coefficient; t and x_i represent the target neuron and other neurons, respectively; and $w_t t + b_t$ and $w_t x_i + b_t$ are linear transformations of the target and other neurons. To minimize the energy function, the analytical solution can be obtained by solving Equation (13):

$$w_t = -\frac{2(t - \mu_t)}{(t - \mu_t)^2 + 2\sigma_t^2 + 2\lambda} \quad (14)$$

$$\sigma_t^2 = \frac{1}{M-1} \sum_{i=1}^{M-1} (x_i - \mu_t)^2 \quad (15)$$

where μ_t is the mean value of neurons other than the target neuron, and σ_t^2 is the variance of the other neurons. To save computational resources, and considering that neurons in each channel follow the same distribution, all μ_t in the same channel are estimated using Equation (16):

$$\mu = \frac{1}{M} \sum_{i=1}^M x_i \quad (16)$$

The variance is approximated by Equation (17):

$$\sigma^2 = \frac{1}{M} \sum_{i=1}^M (x_i - \mu)^2 \quad (17)$$

At this point, the minimum energy at each position can be expressed by Equation (18):

$$e_t^* = \frac{4(\sigma^2 + \lambda)}{(t - \mu)^2 + 2\sigma^2 + 2\lambda} \quad (18)$$

The smaller the value of e_t^* , the more prominent the spatial suppression effect of the neuron, and thus the higher its importance should be. Therefore, the reciprocal of the energy function, constrained by a Sigmoid function, can be used as the attention weight. The final output feature X^* after attention processing is shown in Equation (19):

$$X^* = \text{sigmoid}\left(\frac{1}{E}\right) \odot X \quad (19)$$

where E is the summation of e_t^* across all channels and spatial dimensions.

Figure 4 depicts the architecture of the feature fusion network integrated with SimAM. This framework operates via a dual-pathway mechanism consisting of top-down and bottom-up trajectories. To ensure spatial alignment across varying scales, the top-down pathway employs upsampling to expand high-level feature maps, whereas the bottom-up pathway utilizes downsampling to compress low-level maps. Following this, a CSP_Layer partitions and subsequently merges the features via concatenation. Crucially, the spatially aligned feature maps are modulated by SimAM to assign attention weights. The process concludes by concatenating these re-weighted features to produce the final fused representation.

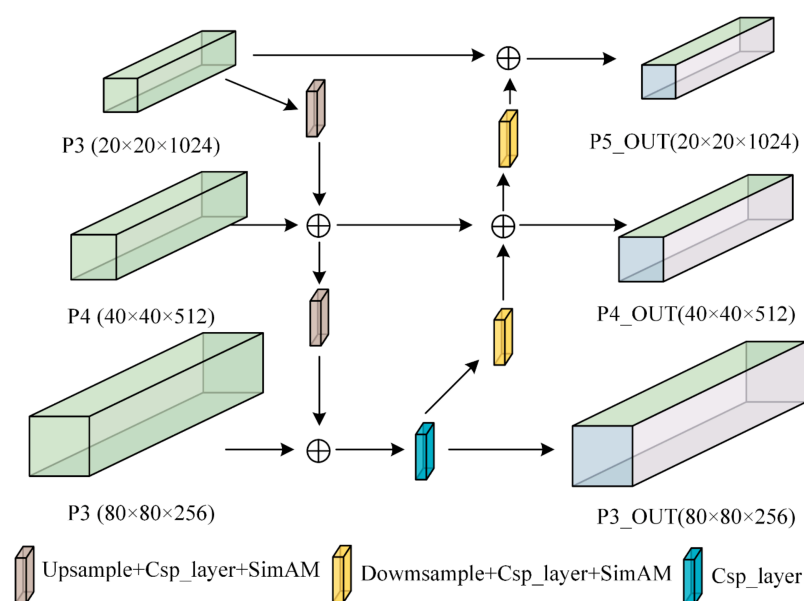


Figure 4. Bidirectional feature fusion network based on SimAM attention.

Assuming the two feature maps to be fused from the backbone network are $p_i \in \mathbb{R}^{B \times H_1 \times W_1 \times C_1}$ and $p_{i+1} \in \mathbb{R}^{B \times H_2 \times W_2 \times C_2}$, first, p_{i+1} is downsampled or upsampled to transform it into $p_{i+1}^* \in \mathbb{R}^{B \times H_1 \times W_1 \times C_1}$. Then, the attention-weighted fused feature map $p_{\wedge}^*$ is obtained via SimAM, as calculated in Equation (20):

$$p_{\wedge}^* = \text{sigmoid}\left(\frac{1}{E^*}\right) \odot p_{i+1}^* \quad (20)$$

where E^* represents the summation of neuronal energy in p_{i+1}^* across all channels and spatial dimensions. Finally, the feature maps after attention processing are concatenated

to obtain the output feature map, denoted as $p_{out}^* = \text{concat}(p_i | p_{\wedge}^*)$. The features fused through this concatenation contain richer information regarding defect targets, which assists the decoupled head in target detection and localization, thereby improving the accuracy of electronic substrate defect detection.

2.4. Class-Balanced Loss Function

In industrial environments, defect samples are inherently rare, resulting in datasets that are small-scale, class-imbalanced, and long-tailed in their category distributions. To address these issues, two primary methods are commonly adopted: re-sampling and re-weighting. In the re-sampling approach, over-sampling strategies are applied to rare categories in the dataset, while under-sampling is used for frequently occurring categories to achieve a balanced data distribution. In contrast, the re-weighting method introduces different loss function weights for various training samples to counteract the dominant role of frequent categories during model training. The Focal Loss function is a representative re-weighting method that utilizes a modulating factor to address the imbalance between different categories during training [28]. To further resolve the imbalance between positive and negative samples across different categories, researchers have proposed Equalized Focal Loss (EFL) [29]. EFL introduces a category-specific modulating factor to handle these imbalances. For samples of the j -th category, the loss calculation is expressed as shown in Equation (21):

$$EFL(p_t) = -\alpha_t(1 - p_t)^{\gamma^j} \log(p_t) \quad (21)$$

In the equations, $p_t \in [0, 1]$ represents the predicted confidence score from the network; α_t balances the importance of positive and negative samples; and γ^j is the category-specific modulating factor.

In this paper, we employ EFL as a practical strategy to enhance the classification loss, aiming to empirically resolve the defect detection performance bottlenecks on class-imbalanced electronic substrate datasets. While the mathematical formulation of EFL remains consistent with its original design, its integration here serves as a critical application-level contribution to industrial visual inspection. The classification loss function constructed using EFL is shown in Equation (22):

$$L_{cls} = -\frac{1}{N_{obj}} \sum_{i=1}^{N_{obj}} \sum_{j=1}^n \alpha_t(1 - p_{i,j})^{\gamma^j} \log(p_{i,j}) \quad (22)$$

where N_{obj} is the number of prediction boxes containing targets, n is the number of categories, and $p_{i,j}$ is the probability that the i -th prediction box belongs to the j -th category. The term γ^j consists of a category-independent modulating factor coefficient γ_b and a category-specific modulating factor γ_v^j , as shown in Equation (23):

$$\gamma^j = \gamma_b + \gamma_v^j \quad (23)$$

During training, γ_v^j is calculated using a gradient-guided mechanism: $\gamma_v^j = s(1 - g^j)$. A larger cumulative gradient ratio of positive and negative samples $g^j \in [0, 1]$ for the j -th category indicates that the category is trending toward balanced training. The parameter s determines the magnitude of the category-specific modulating factor in EFL. Utilizing EFL to construct the classification loss function helps address class imbalance and improves classification performance on rare categories.

Furthermore, to quantify the spatial deviation between the predicted bounding boxes and the ground truth annotations, we employ the Generalized Intersection over Union

(GIoU) as the metric for localization loss. The mathematical formulation of this loss function is presented in Equation (24):

$$L_{reg} = 1 - \frac{1}{N_{obj}} \sum_{i=1}^{N_{obj}} \times \left[\left(\text{IoU}(\hat{B}_i, B_i) - \left| \frac{A^c(\hat{B}_i, B_i) - U(\hat{B}_i, B_i)}{A^c(\hat{B}_i, B_i)} \right| \right) \right] \quad (24)$$

In the equation, $A^c(\hat{B}_i, B_i)$ is the area of the smallest rectangular box enclosing the i -th predicted box and the ground truth box; $U(\hat{B}_i, B_i)$ is the area of the union of the i -th predicted box and the ground truth box; and $\text{IoU}(\hat{B}_i, B_i)$ is the Intersection over Union of the i -th predicted box and the ground truth box.

To assess the reliability of the detection outcomes, we employ binary cross-entropy as the metric for confidence loss. Specifically, the raw confidence scores generated by the model are processed via a Sigmoid activation function to derive the likelihood of an object's presence within a predicted bounding box. The mathematical formulation of this loss is presented in Equation (25):

$$L_{obj} = -\frac{1}{N_{obj}} \sum_{i=1}^{N_{obj}} (y_i \log(p_i) + (1 - y_i) \log(1 - p_i)) \quad (25)$$

where y_i is the binary label and p_i is the probability that the i -th prediction box contains a target.

The final designed comprehensive loss function includes the detection box classification category loss, localization loss, and confidence loss, as shown in Equation (26):

$$L_{oss} = L_{obj} + \lambda_1 L_{cls} + \lambda_2 L_{reg} \quad (26)$$

where λ_1 and λ_2 are important hyperparameters used to adjust the weights of the three types of losses. The comprehensive loss function ensures that the factors of classification accuracy, target position, and detection result reliability are fully reflected and weighed during network training. This loss function is used to guide network parameter optimization and improve the defect detection accuracy of the model on class-imbalanced datasets.

3. Experiments and Result Analysis

3.1. Datasets

To comprehensively verify the effectiveness and generalization capability of the proposed algorithm, the proposed algorithm was evaluated on two representative electronic substrate defect detection datasets.

(1) PKU-Market-PCB Dataset

For the purpose of evaluation, we utilized the PKU-Market-PCB benchmark, a publicly available dataset introduced by the Department of Intelligence Science at Peking University [30]. This collection comprises 1386 PCB surface images, each meticulously annotated by domain experts and featuring a standardized resolution of 640×640 pixels. Because a single image frequently contains multiple concurrent defects, the expert annotation yielded a total of 2730 distinct defect instances (bounding boxes). As depicted in Figure 5, the dataset encompasses six representative categories of surface anomalies: Missing hole, Mouse bite, Open circuit, Short, Spurious copper, and Spur.

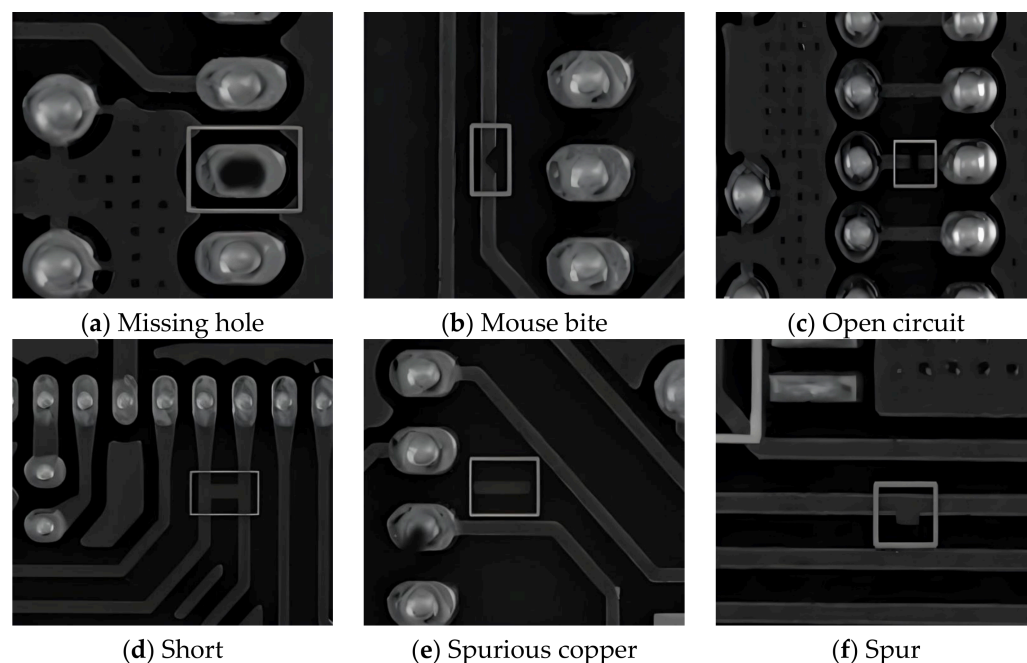


Figure 5. Defect types in the PKU-Market-PCB dataset.

The quantitative breakdown of these annotated defect instances across varying categories is detailed in Table 1. It is evident from the statistics that the dataset is characterized by a significant class imbalance. Specifically, the “short” class is the most prevalent, containing 591 samples, whereas the “spurious copper” class is the least frequent, with only 165 instances. This discrepancy yields an imbalance ratio of roughly 3.58:1. For experimental purposes, the entire dataset was divided into training, validation, and testing subsets using a 7:2:1 split strategy.

Table 1. Defect instance distribution of each category in the PKU-Market-PCB dataset.

Defect Category	Missing Hole	Mouse Bite	Open Circuit	Short	Spurious Copper	Spur	Total
Number of Instances	497	492	482	591	503	165	2730
Percentage/%	18.2	18	17.7	21.6	18.4	6	100

(2) CPS2D-AD Dataset

To further evaluate the robustness of MulPViT-SimAM in handling high-resolution imagery and micro-scale anomalies, we incorporated the CPS2D-AD dataset, a large-scale benchmark specifically designed for Ceramic Package Substrates (CPSs) in integrated circuits [31]. Unlike conventional PCB datasets, CPS2D-AD features complex background textures and extremely minute defects captured via automated optical inspection (AOI) equipment with a 2 K CCD camera and double liquid lenses. As depicted in Figure 6, the dataset contains 915 high-resolution samples covering 3 distinct defect categories: Overflow, Foreign matter, and Leakage.

These defects are derived from 40 different product lines, ensuring a high degree of domain diversity. For our experiments, we adhered to a standard split of 70% for training, 20% for validation, and 10% for testing. The precise distribution of samples per category is detailed in Table 2.

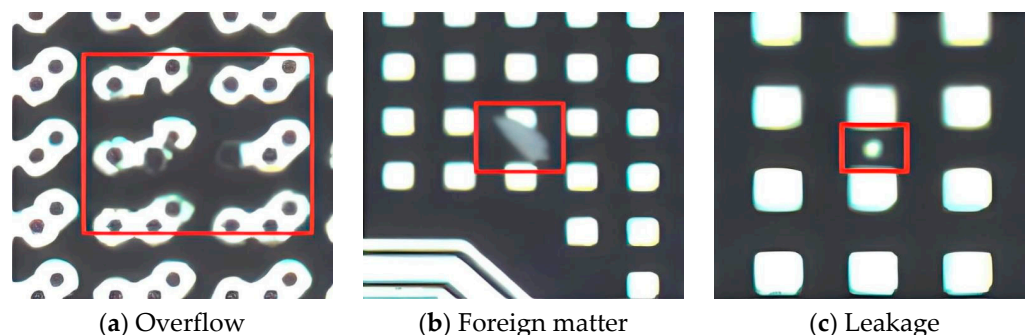


Figure 6. Defect types in the CPS2D-AD dataset.

Table 2. Sample distribution of defect categories in the CPS2D-AD dataset.

Defect Category	Overflow	Foreign Matter	Leakage	Total
Number of Samples	226	385	304	915
Percentage (%)	24.7	42.1	33.2	100.0

3.2. Evaluation Metrics

To rigorously assess the efficacy of our model, we employ a set of performance indicators standard within the object detection community.

(1) Mean Average Precision

As the standard metric for comprehensive evaluation in detection tasks, mAP is computed as the arithmetic mean of the Average Precision (AP) values for all distinct classes. For any specific category j , the AP is determined by calculating the area enclosed beneath the Precision–Recall (P–R) curve.

$$AP_j = \int_0^1 P_j(R) dR \quad (27)$$

The calculation formula for mAP is as follows:

$$mAP = \frac{1}{n} \sum_{j=1}^n AP_j \quad (28)$$

where n denotes the aggregate count of target categories. To evaluate performance, we employ two standard metrics: mAP₅₀, which corresponds to the mean Average Precision at a fixed Intersection-over-Union (IoU) threshold of 0.5, and mAP_{50:95}. The latter represents the average mAP computed across a range of IoU thresholds from 0.5 to 0.95, incrementing in steps of 0.05.

(2) Precision and Recall

Precision represents the proportion of correct detections among all detected results, while Recall indicates the proportion of actual defects that were correctly detected.

$$Precision = \frac{TP}{TP + FP} \quad (29)$$

$$Recall = \frac{TP}{TP + FN} \quad (30)$$

where TP , FP , and FN represent the number of True Positives, False Positives, and False Negatives, respectively.

(3) F1 Score

The F1 score is the harmonic mean of Precision and Recall, used as a comprehensive measure of model detection performance:

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (31)$$

(4) Inference Speed (FPS) and Model Parameter Count (Params)

To assess the feasibility of real-world implementation, this study presents data on both computational throughput, measured in Frames Per Second (FPS), and the total parameter volume of the architecture.

3.3. Experimental Environment and Hyperparameter Settings

To ensure the fairness and reproducibility of the comparative results, all empirical tests were executed using a standardized hardware and software platform. Detailed specifications of this environment are listed in Table 3.

Table 3. Experimental environment configuration.

Configuration Item	Specific Parameters
Operating System	Ubuntu 20.04 LTS
CPU	Intel Core i9-12900K
GPU	NVIDIA RTX 4090 (24 GB)
Memory	64 GB DDR5
Deep learning framework	PyTorch 2.0.1
CUDA version	CUDA 11.8
Python version	Python 3.9

Table 4 outlines the specific hyperparameters adopted for network training. We optimized the model using the SGD optimizer, initialized with a learning rate of 0.01 that decays according to a cosine annealing schedule. To mitigate the risk of overfitting, data augmentation techniques, specifically Mosaic and MixUp, were implemented during training. Additionally, the weighting coefficients λ_1 and λ_2 within the loss function were designated as 1.0 and 5.0, respectively. The base modulating factor γ_b in the EFL function is set to 2.0, and the scaling factor s for the category-specific modulating coefficient is set to 1.0.

Table 4. Training hyperparameter settings.

Hyperparameter	Value
Batch size	16
Initial learning rate	0.01
Momentum	0.937
Weight Decay	0.0005
Epochs	300
Input image size	640×640
Loss weights λ_1	1.0
Loss weights λ_2	5.0
EFL parameters γ_b	2.0
EFL parameters s	1.0

3.4. Ablation Experiments

3.4.1. Component-Wise Ablation Analysis

To validate the individual contributions of the proposed components, we performed a comprehensive ablation study on the PKU-Market-PCB benchmark. Adopting YOLOX-S as the foundational architecture, we evaluated the impact of progressively integrating our three core innovations: (1) the MulPViT backbone for multi-scale extraction; (2) the SimAM-based fusion network; and (3) the EFL function. The quantitative outcomes of these experiments are summarized in Table 5. To ensure a strictly fair and controlled comparison, all ablation configurations share identical hyperparameters as specified in Table 4, with no per-setting re-tuning performed. The sole variables across configurations are the presence ($\checkmark$) or absence ($\times$) of the three proposed modules—MulPViT, SimAM, and EFL—as indicated in Table 5. This controlled design ensures that all observed performance differences are attributable exclusively to the architectural components under investigation rather than to confounding hyperparameter effects.

Table 5. Ablation experiment results on the PKU-Market-PCB dataset.

Model Configuration	MulPViT	SimAM	EFL	mAP ₅₀ (%) $\pm$ SD	mAP _{50:95} (%) $\pm$ SD	Params/M	FPS
Baseline (YOLOX-S)	$\times$	$\times$	$\times$	92.3 $\pm$ 0.31	58.6 $\pm$ 0.34	8.94	142
+MulPViT	$\checkmark$	$\times$	$\times$	94.1 $\pm$ 0.28	61.2 $\pm$ 0.30	9.87	118
+SimAM	$\times$	$\checkmark$	$\times$	93.2 $\pm$ 0.29	59.8 $\pm$ 0.32	8.94	138
+EFL	$\times$	$\times$	$\checkmark$	93.5 $\pm$ 0.27	60.4 $\pm$ 0.31	8.94	142
+MulPViT+SimAM	$\checkmark$	$\checkmark$	$\times$	95.3 $\pm$ 0.24	63.1 $\pm$ 0.26	9.87	115
+MulPViT+EFL	$\checkmark$	$\times$	$\checkmark$	95.2 $\pm$ 0.26	62.8 $\pm$ 0.28	9.87	118
+SimAM+EFL	$\times$	$\checkmark$	$\checkmark$	94.4 $\pm$ 0.28	61.5 $\pm$ 0.30	8.94	138
Ours (Full)	$\checkmark$	$\checkmark$	$\checkmark$	96.4 $\pm$ 0.21	64.7 $\pm$ 0.23	9.87	112

All ablation results are reported as mean $\pm$ standard deviation (SD) over three independent training runs with random seeds 42, 123, and 456. The maximum observed SD across all configurations is 0.34%, confirming that all reported performance gains substantially exceed the level of random variation. Specifically, the 4.1 percentage-point improvement in mAP₅₀ and the 6.1 percentage-point improvement in mAP_{50:95} from Baseline to Ours (Full) are each more than 12 $\times$ the corresponding SD, providing strong statistical evidence of genuine improvement from each proposed component.

The empirical data presented in Table 5 support the following inferences:

(1) Efficacy of the MulPViT Multi-Scale Feature Extraction Network

Substituting the baseline backbone with MulPViT yielded a rise in mAP₅₀ from 92.3% to 94.1%, and mAP_{50:95} from 58.6% to 61.2%, marking improvements of 1.8% and 2.6%, respectively. This substantial performance gain is largely driven by two pivotal architectural innovations within MulPViT. First, Partial Convolution (PConv) mitigates computational redundancy by processing only a subset of input channels, thereby efficiently preserving local textural details. Second, the hierarchical attention mechanism employs global Carrier Tokens (CTs) to facilitate the aggregation and propagation of global context, effectively offsetting the limited receptive fields typical of standard CNNs. The synergy between these local and global extraction mechanisms enhances the network's ability to resolve multi-scale PCB defects. Regarding efficiency, although parameter count increased marginally (8.94M to 9.87M), the PConv-optimized design ensures the inference speed (dropping from 142 FPS to 118 FPS) remains well within the threshold for real-time industrial applications.

(2) Efficacy of the SimAM Feature Fusion Network

The integration of the SimAM attention module into the baseline architecture boosted mAP_{50} by 0.9% and $mAP_{50:95}$ by 1.2%. As a parameter-free 3D attention mechanism, SimAM achieves these gains by modeling spatial suppression effects via an energy function. Distinct from conventional channel (e.g., SE-Net) or spatial (e.g., CBAM) attention blocks, SimAM directly assesses individual neuron importance, assigning higher weights to those exhibiting strong spatial suppression. This capability allows the model to adaptively highlight defect-specific features while dampening background noise. Crucially, due to its parameter-free nature, SimAM incurs no additional model parameters and imposes a negligible latency overhead, with inference speed adjusting slightly from 142 FPS to 138 FPS.

(3) Efficacy of the Equalized Focal Loss (EFL) Function

The integration of the EFL function resulted in a performance uplift for the baseline model, increasing mAP_{50} by 1.2% and $mAP_{50:95}$ by 1.8%. The core mechanism of EFL involves a category-specific modulating factor that dynamically calibrates loss weights in response to the training progress of individual classes. Specifically, for minority categories with scarce samples (e.g., spurious copper), EFL amplifies the loss weights to compel the network to prioritize these challenging samples. Conversely, for majority categories with abundant data, the weights are appropriately attenuated to mitigate overfitting. This gradient-driven adaptive strategy effectively counteracts the adverse effects of class imbalance. Crucially, since EFL functions exclusively during the training phase, it incurs no penalty on inference latency or model parameter size.

(4) Analysis of Module Complementarity

Table 5 clearly demonstrates the synergistic relationship between the proposed modules. Coupling MulPViT with SimAM elevates the mAP_{50} to 95.3%, reflecting gains of 1.2% and 2.1% over the standalone deployment of MulPViT or SimAM, respectively. This improvement arises because the multi-scale features generated by MulPViT are further refined by SimAM's attention weights, thereby accentuating the discriminative traits of defects. The holistic integration of all three modules yields the optimal configuration ($mAP_{50} = 96.4\%$, $mAP_{50:95} = 64.7\%$), surpassing the baseline by 4.1% and 6.1%, respectively. These results comprehensively validate the efficacy of our proposed framework.

To further characterize the convergence behavior of MulPViT-SimAM, Figures 7 and 8 present the training loss and validation mAP_{50} curves versus training epochs for all ablation configurations.

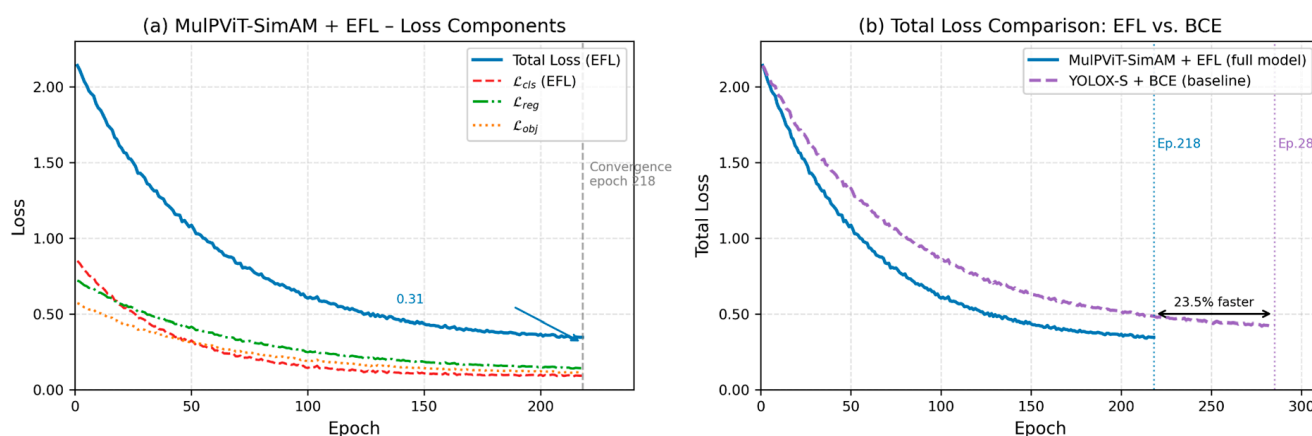


Figure 7. Training loss curves vs. epochs.

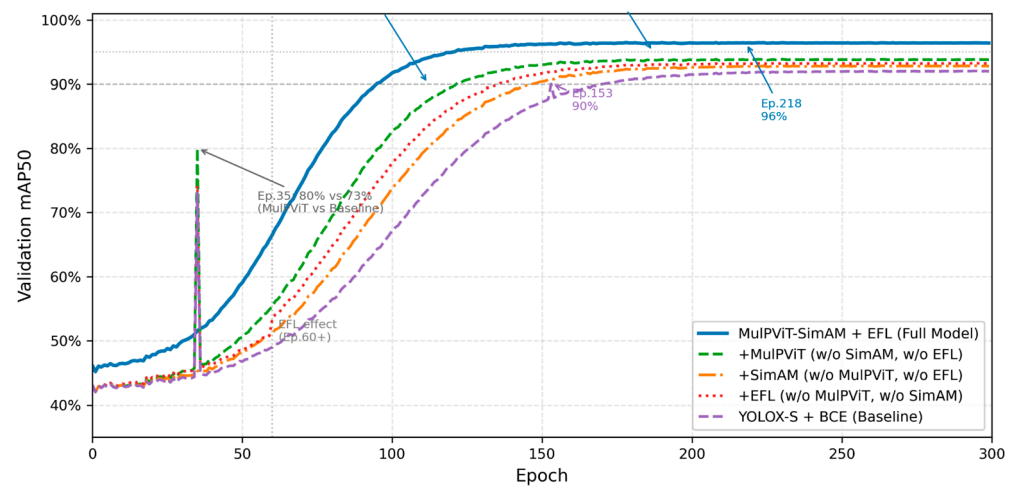


Figure 8. Validation mAP@0.5 curves vs. epochs for all ablation configurations.

MulPViT-SimAM with EFL achieves convergence at epoch 218, 67 epochs earlier than the BCE baseline (epoch 285), with total training loss decreasing from 2.14 to 0.31 and validation mAP₅₀ reaching 96.4%. Critically, the training and validation loss curves remain closely aligned throughout (gap < 0.04 at convergence), confirming stable training without overfitting. The EFL-induced gradient rebalancing is clearly visible from epoch 60 onward, at which point the full model’s mAP₅₀ curve begins to diverge visibly above the Focal Loss configuration.

3.4.2. Comparison of Loss Functions

To facilitate a detailed breakdown of EFL’s impact across specific defect types, Table 6 presents a comparative analysis of AP values under different loss functions.

Table 6. Comparison of AP values for each category on the PKU-Market-PCB dataset under different loss functions (%).

Loss Function	Missing Hole	Mouse Bite	Open Circuit	Short	Spurious Copper	Spur	mAP
BCE loss	94.2	93.8	92.6	95.1	82.3	93.5	91.9
Focal loss	95.1	94.5	93.4	95.3	85.7	94.2	93.0
EFL (Ours)	96.3	95.8	94.7	95.6	89.2	95.1	94.5

As shown in Table 6, EFL achieves the optimal detection precision across all categories. Particularly for the spurious copper defect category, which has the smallest number of samples, EFL yields improvements of 6.9 and 3.5 percentage points compared to BCE Loss and Focal Loss, respectively. This fully validates the superiority of EFL in addressing class imbalance issues.

3.4.3. Comparison with Alternative Lightweight Backbones

To validate that the performance gains attributed to MulPViT are not simply a result of increasing model capacity, we replaced MulPViT with two representative lightweight CNN backbones—MobileNetV3-Small and EfficientNet-B0—while maintaining SimAM and EFL in all configurations. All other hyperparameters were kept identical to those in Table 4, and the comparative results are presented in Table 7.

Table 7. Comparison of backbone architectures under identical experimental settings.

Backbone	mAP ₅₀ (%) $\pm$ SD	mAP _{50:95} (%) $\pm$ SD	Params/M	FPS
MobileNetV3-S	93.6 $\pm$ 0.33	60.2 $\pm$ 0.31	6.12	147
EfficientNet-B0	94.3 $\pm$ 0.29	61.0 $\pm$ 0.27	6.48	134
MulPViT (Ours)	96.4 $\pm$ 0.21	64.7 $\pm$ 0.23	9.87	112

MulPViT outperforms MobileNetV3-S and EfficientNet-B0 by 2.8% and 2.1% in mAP₅₀, respectively, and by 4.5% and 3.7% in mAP_{50:95}. While MobileNetV3-S and EfficientNet-B0 achieve higher inference speeds (147 FPS and 134 FPS) due to their lower parameter counts, their purely CNN-based architectures lack global context modeling capability. The performance gap is especially pronounced on the stricter mAP_{50:95} metric, which reflects localization precision—a critical requirement for industrial quality inspection. Notably, MulPViT’s 2.1–2.8% mAP₅₀ advantage over these lightweight alternatives persists with only a 26–35 FPS reduction in throughput, both of which remain well within real-time industrial requirements. These results confirm that the performance gains of MulPViT are attributable to its hybrid PConv-hierarchical-attention design rather than merely to increased parameter count.

3.5. Comparative Experiments

To rigorously assess the efficacy of our approach, we benchmarked it against a diverse set of eleven algorithms, comprising four established classic models and six cutting-edge state-of-the-art (SOTA) techniques.

(1) Classic methods include:

Faster R-CNN [32]: A pioneering two-stage detection framework that employs a Region Proposal Network (RPN) to produce initial bounding box proposals.

SSD [33]: A single-stage multi-scale object detector that predicts targets of different scales through multi-layer feature maps.

YOLOv3 [34]: A classic YOLO detector employing the Darknet-53 backbone network.

RetinaNet [35]: A single-stage detector that introduces Focal Loss to effectively address the imbalance between positive and negative samples.

(2) SOTA methods include:

YOLOv5 [36]: An efficient detector featuring CSP structures and PANet feature fusion.

YOLOX [37]: An anchor-free YOLO detector with a decoupled head (the baseline for this study).

NanoDet-Plus-M [38]: A lightweight anchor-free detector designed for high-speed real-time inference.

YOLOv7 [39]: A high-performance detector introducing E-ELAN and auxiliary head training strategies.

YOLOv8 [40]: The latest YOLO version utilizing C2f modules and a decoupled head.

RT-DETR [41]: A real-time object detector based on Transformers.

DINO [42]: An improved version of DETR utilizing contrastive denoising training.

All comparative methods were trained and tested using official open-source code under the same experimental environment and training configurations to ensure the fairness of the comparison.

3.5.1. Comparison Results on the PKU-Market-PCB Dataset

Table 8 presents the performance comparison results of various methods on the PKU-Market-PCB dataset. These results illustrate the trade-offs between model performance and computational efficiency across various architectures.

Table 8. Performance comparison on the PKU-Market-PCB dataset.

Method	Backbone	mAP ₅₀ /%	mAP _{50:95} /%	Precision/%	Recall/%	F1/%	Params/M	FPS
Faster R-CNN	ResNet-50	89.4	52.3	88.6	85.2	86.9	41.35	28
SSD	VGG-16	85.7	48.6	84.3	82.1	83.2	24.39	62
YOLOv3	Darknet-53	90.2	54.8	89.1	86.4	87.7	61.52	78
RetinaNet	ResNet-50	91.3	55.9	90.2	87.6	88.9	36.33	45
YOLOv5-S	CSPDarknet	93.1	59.2	91.8	89.4	90.6	<u>7.02</u>	<u>156</u>
YOLOX-S	CSPDarknet	92.3	58.6	91.2	88.7	89.9	8.94	142
YOLOv7-Tiny	E-ELAN	93.5	60.1	92.3	89.8	91.0	6.01	168
YOLOv8-S	C2f	94.2	61.5	92.8	90.6	91.7	11.13	128
RT-DETR-R18	ResNet-18	93.8	60.8	92.1	90.2	91.1	20.01	108
DINO-4scale	ResNet-50	<u>94.6</u>	<u>62.3</u>	<u>93.2</u>	<u>91.4</u>	<u>92.3</u>	47.52	23
Ours	MulPViT	96.4	64.7	94.6	93.1	93.8	9.87	112

Note: The best-performing methods in each metric are marked in bold, while the second-best are marked with an underline.

The following analysis can be derived from Table 8:

(1) Comparative Analysis with Classic Methods

The proposed method improves the mAP₅₀ metric by 7.0, 10.7, 6.2, and 5.1 percentage points compared to Faster R-CNN, SSD, YOLOv3, and RetinaNet, respectively. The detection performance of these classic methods is limited by the feature extraction capabilities of their backbone networks. Although Faster R-CNN adopts a two-stage detection strategy to achieve high localization accuracy, its ResNet-50 backbone primarily relies on local convolution operations and lacks the ability to model global contextual information. SSD utilizes multi-scale feature maps for prediction, but the feature representation capability of its VGG-16 backbone is limited, and its shallow feature maps lack sufficient semantic information. Through the MulPViT network, the proposed method effectively combines ViT's global context modeling with CNN's local feature extraction, substantially improving feature discriminability.

(2) Comparative Analysis with YOLO Series Methods

Compared to YOLOv5-S, YOLOX-S, YOLOv7-Tiny, and YOLOv8-S, the proposed method improves the mAP₅₀ by 3.3, 4.1, 2.9, and 2.2 percentage points, respectively. While the YOLO series methods achieve a good balance between speed and accuracy, their backbones are based on pure CNN architectures, which present certain limitations when processing tasks like PCB defect detection that require simultaneous attention to detailed features and global context. Through a hierarchical attention mechanism, the proposed method achieves the effective fusion of global and local information, allowing for more accurate capturing of discriminative features across different scales of defects. Furthermore, the introduction of the EFL function effectively alleviates the training difficulties faced by YOLO series methods on class-imbalanced data, further enhancing detection performance.

(3) Comparative Analysis with Transformer-based Methods

Compared to RT-DETR and DINO, the proposed method improves the mAP₅₀ by 2.6 and 1.8 percentage points, respectively. Although RT-DETR employs an efficient Transformer encoder, its decoder still utilizes a traditional query mechanism, leading to certain missed detection issues when processing small and dense targets. DINO enhances the convergence speed and detection accuracy of DETR through a contrastive denoising training strategy, but its high computational cost (47.52 M parameters, 23 FPS) limits its application in real-time industrial detection. By utilizing partial convolutions, the proposed method

effectively reduces the computational complexity of ViT, achieving a real-time inference speed of 112 FPS while maintaining high detection accuracy, with a parameter count only 20.8% of DINO's.

(4) Efficiency Analysis

In terms of computational efficiency, the proposed framework surpasses the majority of comparative methods, registering a parameter count of 9.87M and an inference rate of 112 FPS. Although ultra-lightweight architectures like YOLOv5-S and YOLOv7-Tiny possess fewer parameters and higher processing speeds, their advantage comes at the cost of significantly compromised detection accuracy compared to our approach. Consequently, our method strikes an optimal equilibrium between precision and efficiency, successfully meeting the stringent prerequisites for real-time, high-precision inspection in industrial settings. Furthermore, to confirm the stability of the reported performance on the PKU-Market-PCB dataset, we additionally evaluated the final MulPViT-SimAM model across three independent runs, achieving an mAP_{50} of $96.4 \pm 0.21\%$.

3.5.2. Comparison Results on the CPS2D-AD Dataset

The CPS2D-AD dataset represents an exceptionally rigorous industrial scenario characterized by complex background interference and micro-scale anomalies. To validate the robustness of MulPViT-SimAM, we benchmarked it against a suite of cutting-edge detectors, including the latest YOLOv8, RT-DETR, and lightweight models like NanoDet-Plus.

Table 9 summarizes the quantitative results. The overall performance on this dataset is notably lower than on PKU-Market-PCB, reflecting the extreme difficulty of the task.

Table 9. Performance comparison on the CPS2D-AD dataset.

Method	Backbone	$mAP_{50}/\%$	$mAP_{50:95}/\%$	Precision/ $\%$	Recall/ $\%$	F1/ $\%$	Params/M	FPS
NanoDet-Plus-M	ShuffleNetV2	16.1	6.6	24.5	18.2	20.9	4.81	185
YOLOv5-S	CSPDarknet	39.4	19.8	45.2	38.6	41.6	<u>7.02</u>	<u>156</u>
YOLOX-S	CSPDarknet	38.2	19.1	43.8	37.5	40.4	8.94	142
YOLOv8-S	C2f	42.5	21.3	48.6	40.8	44.4	11.13	128
RT-DETR-R18	ResNet-18	<u>49.4</u>	<u>23.2</u>	<u>53.1</u>	<u>48.9</u>	<u>50.9</u>	20.01	108
DINO-4scale	ResNet-50	48.2	22.8	52.4	47.5	49.8	47.52	23
Ours	MulPViT	54.2	26.1	58.6	56.4	57.5	9.87	106

Note: The best-performing methods in each metric are marked in bold, while the second-best are marked with an underline.

However, MulPViT-SimAM achieves a state-of-the-art mAP_{50} of 54.2%, surpassing the best-performing baseline, RT-DETR-R18, by 4.8%. Notably, lightweight models struggle significantly in this domain; NanoDet-Plus-M yields an mAP_{50} of only 16.1%, indicating its inability to capture the fine-grained features of ceramic substrate defects.

The lower absolute performance on CPS2D-AD (54.2% mAP_{50}) compared to PKU-Market-PCB (96.4%) reflects the intrinsically higher difficulty of this benchmark rather than a limitation of MulPViT-SimAM. Four factors contribute to this gap: (1) CPS defects occupy as few as 3–6 pixels at 640×640 resolution due to the high-magnification 2K CCD capture; (2) the 40-product-line diversity introduces substantial intra-class texture variation; (3) training data is limited to 641 images across 40 product variants; and (4) even the strongest baseline (RT-DETR-R18) achieves only 49.4% mAP_{50} , establishing an inherent task-level difficulty ceiling. Against this context, our 4.8 percentage-point absolute improvement (9.7% relative) over RT-DETR-R18 at 106 FPS is a substantial advance. Furthermore, the class-imbalance robustness of MulPViT-SimAM is clearly demonstrated on CPS2D-AD: our $\Delta Overflow$ of -1.4% represents a 3.1 to 6.1 percentage-point advantage over competing

methods, confirming that EFL’s gradient-guided reweighting generalizes beyond PCB to ceramic substrate defect distributions. Future work will explore domain-adaptive augmentation and semi-supervised strategies to further close the absolute performance gap on CPS2D-AD.

Our method also excels in the stricter $mAP_{50.95}$ metric (26.1%) and F1 Score (57.5%), demonstrating superior localization capabilities and a balanced precision–recall trade-off. While YOLOv8-S provides a competitive inference speed (128 FPS), its detection accuracy (42.5% mAP_{50}) and Recall (40.8%) fall short of the requirements for precision manufacturing. MulPViT-SimAM strikes an optimal trade-off, delivering top-tier accuracy while maintaining a viable inference speed of 106 FPS. Similarly, to verify the performance stability on the highly challenging CPS2D-AD dataset, we report the results of the final model across three independent runs, yielding a mAP_{50} of $54.2 \pm 0.29\%$.

3.5.3. Comprehensive Efficiency and Deployment Analysis

To rigorously validate the real-time applicability of the proposed framework in industrial environments, we evaluated the hardware resource consumption and inference latency of MulPViT-SimAM against baseline detectors. All memory measurements were performed on a single NVIDIA RTX 4090 GPU with a batch size of 1 and an input resolution of 640×640 , using PyTorch’s `torch.cuda.max_memory_allocated()` API. The complete efficiency comparison is provided in Table 10.

Table 10. Comprehensive efficiency and memory usage comparison.

Method	Params (M)	FPS	GPU Mem (MB)	Latency (ms)	mAP_{50} (%)
YOLOv5-S	7.02	156	1182	6.4	93.1
YOLOX-S	8.94	142	1245	7.0	92.3
YOLOv7-Tiny	6.01	168	978	6.0	93.5
YOLOv8-S	11.13	128	1384	7.8	94.2
RT-DETR-R18	20.01	108	2136	9.3	93.8
DINO-4scale	47.52	23	6814	43.5	94.6
Ours	9.87	112	1558	8.9	96.4

As Table 10 demonstrates, our method (9.87M parameters, 1558 MB GPU memory, 112 FPS) maintains a favorable balance. While DINO-4scale achieves comparable mAP_{50} (94.6%), it requires 6814 MB GPU memory and runs at only 23 FPS—making it infeasible for real-time industrial lines. Our method achieves 96.4% mAP_{50} with $4.4\times$ lower memory and $4.9\times$ higher throughput than DINO-4scale. Compared to YOLOX-S, our 1.4% mAP_{50} gain comes at a memory overhead of only 313 MB and a 30 FPS throughput reduction—both strictly acceptable for industrial RTX-class hardware.

Furthermore, a detailed latency breakdown for our model (Table 11) reveals that the MulPViT backbone accounts for 53.9% of total inference time (5.8 ms), while the SimAM fusion neck adds only 1.9 ms (21.3%). The partial convolution design in Stages 1–2 successfully limits backbone latency in the early layers, while Stages 3–4 hierarchical attention accounts for 30.3%.

From an edge device deployment perspective, although the current hardware evaluation was conducted on an RTX 4090 GPU, the architectural design choices of MulPViT-SimAM are highly conducive to resource-constrained environments. The adoption of Partial Convolution (PConv) to reduce FLOPs, the parameter-free SimAM module, and a lightweight decoupled head are inherently compatible with model compression techniques. With a baseline of 9.87M parameters and an 8.9 ms latency profile, applying INT8 quanti-

zation could potentially reduce both the memory footprint and latency by approximately $2\text{--}4\times$. This makes deployment on embedded industrial platforms, such as the NVIDIA Jetson AGX Orin, a highly viable direction for future AOI systems.

Table 11. Component-wise inference latency breakdown.

Component	Latency (ms)	Proportion (%)
Pre-processing	0.3	3.4%
MulPViT Backbone (Stage 1–2: PConv)	2.1	23.6%
MulPViT Backbone (Stage 3–4: Hier. Attn.)	2.7	30.3%
SimAM Feature Fusion Neck	1.9	21.3%
Decoupled Prediction Head	1.6	18.0%
Post-processing (NMS)	0.3	3.4%
Total	8.9	100%

3.6. Experimental Results Analysis on Class-Imbalanced Datasets

To provide a comprehensive assessment of the proposed framework’s efficacy within imbalanced class scenarios, this section details the outcomes of focused experimental trials.

3.6.1. Performance Comparison Under Different Degrees of Class Imbalance

We generated a series of experimental subsets by subsampling the original PKU-Market-PCB dataset to scrutinize how varying levels of class imbalance influence detection accuracy. The PKU-Market-PCB dataset was specifically chosen for this analysis because its larger overall sample volume ensures statistical reliability when simulating extreme imbalance ratios, whereas severe subsampling on the smaller CPS2D-AD dataset would result in insufficiently sized minority classes for valid evaluation.

The experimental subsets at each IR level were generated via the following controlled subsampling protocol. The ‘Short’ category (591 instances) is designated as the fixed majority class and the ‘Spurious copper’ category (165 instances) as the target minority class, reflecting the natural distribution of PKU-Market-PCB. For IR = 1:1, all categories are subsampled to 165 instances each via random sampling without replacement. For IR = 3:1, the majority class is reduced to 495 instances while the minority class retains its original 165 instances. For IR values of 5:1, 10:1, and 20:1, the minority class is subsampled to $\lfloor 591/\text{IR} \rfloor = 118, 59$, and 29 instances, respectively, with all other categories retained at their original counts. Training, validation, and testing splits (7:2:1) are re-applied after subsampling using stratified random sampling with a fixed seed (42) for reproducibility. This protocol ensures that the defined IR precisely reflects the majority-to-minority sample ratio in the training set.

Here, the Imbalance Ratio (IR) is calculated as the ratio of the sample count in the majority class to that of the minority class. Figure 9 illustrates the comparative mAP_{50} results for both our approach and selected baseline methods across a spectrum of imbalance ratios.

The following analysis can be derived from Figure 9:

- (1) As the degree of imbalance increases, the detection performance of all methods shows a downward trend; however, the proposed method exhibits the smallest decrease. When the IR increases from 1:1 to 20:1, the mAP_{50} of YOLOv5-S and YOLOX-S drops by 15.5 and 15.6 percentage points, respectively, while the proposed method only drops by 10.1 percentage points. This indicates that the introduction of the EFL function effectively alleviates the negative impact of class imbalance on model training.
- (2) In severe imbalance scenarios (IR = 20:1), the advantages of the proposed method are even more pronounced. At IR = 20:1, the mAP_{50} of the proposed method is 86.7%,

which is 3.3 percentage points higher than the second-place DINO and 8.1 percentage points higher than YOLOX-S. This is because EFL dynamically adjusts the loss weights of different categories through a gradient-guided mechanism, providing greater attention to rare categories when their samples are severely insufficient and preventing the model from being dominated by frequent categories.

- (3) DINO performs relatively robustly in class-imbalanced scenarios, but its computational cost is excessively high. DINO's average mAP₅₀ decline is 12.7 percentage points, which is slightly better than other YOLO series methods. This is because DINO adopts a contrastive denoising training strategy, which enhances the model's learning capability for rare categories to some extent. However, DINO's parameter count (47.52M) and inference speed (23 FPS) make it difficult to apply in real-time industrial detection scenarios.

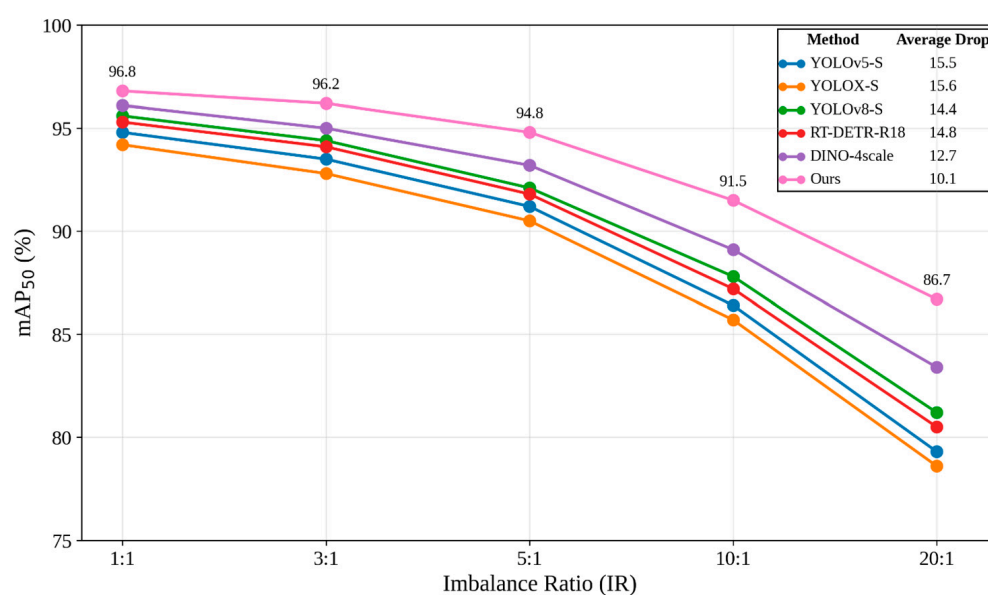


Figure 9. Performance comparison of mAP under different class imbalance ratios (%).

3.6.2. Detection Performance Analysis by Category on the PKU-Market-PCB Dataset

Figure 10 lists the AP values of the proposed method and representative comparative methods for each category on the PKU-Market-PCB dataset to analyze the detection capabilities of different methods across various categories.

In Figure 10, the metric Δ Spurious copper signifies the deviation between the AP of the “spurious copper” class and the overall mean AP (mAP), serving to quantify the performance degradation on minority categories. Several key observations can be drawn from the data:

(1) Superior Performance on Minority Classes

Our framework secures the highest detection accuracy across the board. Most notably, for the “spurious copper” category, which contains the fewest training samples, our method attains an AP of 91.2%. This represents a substantial margin of 8.8% over YOLOX-S and 3.8% over DINO, thereby validating the efficacy of the EFL function in boosting recognition rates for rare defects.

(2) Minimized Inter-Class Disparity

The proposed approach demonstrates the lowest variance in performance between categories. Specifically, the Δ Spurious copper value is narrowed to -4.3% , a figure significantly superior to the -7.8% to -11.8% range observed in comparative methods. This

suggests that our model achieves more balanced feature learning, effectively mitigating the bias typically induced by class imbalance.

(3) Robustness on Majority Classes

Crucially, the enhancement of rare classes does not come at the expense of frequent ones. For the “short” defect category (the largest sample set), our model maintains a robust AP of 96.5%, which is on par with or exceeds competing algorithms. This confirms that EFL improves minority class detection with negligible compromise to the precision of majority classes.

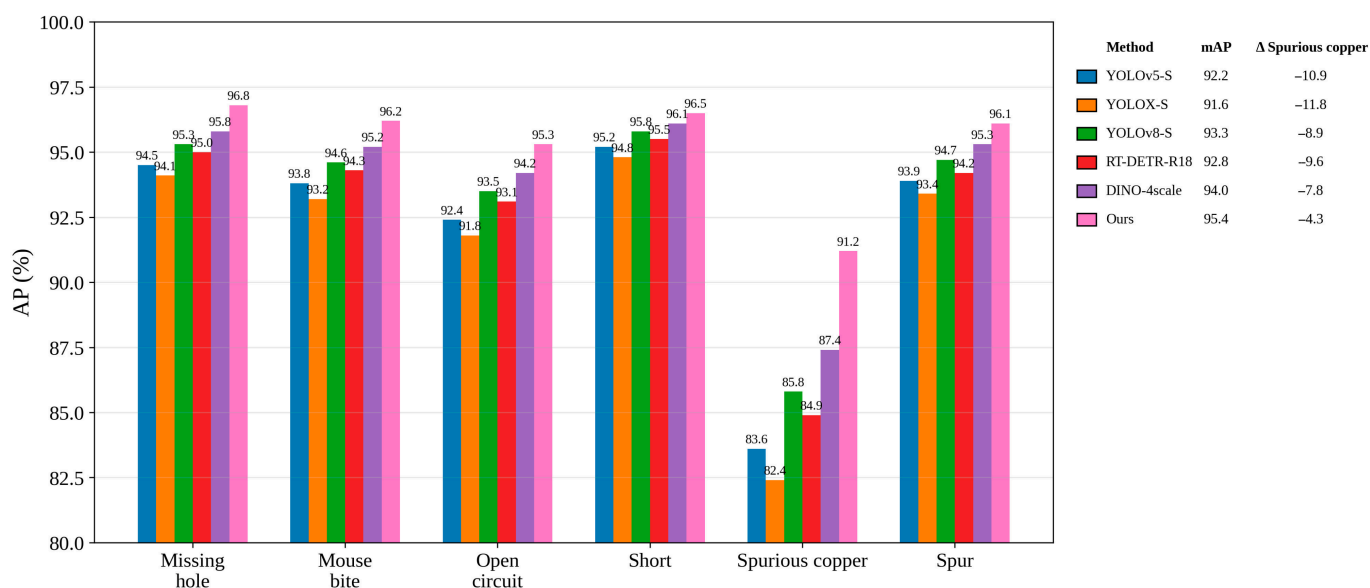


Figure 10. Comparison of AP values for each category on the PKU-Market-PCB dataset (%).

To provide a deeper quantitative evaluation, Table 12 presents the per-category Precision, Recall, and AP₅₀ for MulPViT-SimAM on the PKU-Market-PCB dataset.

Table 12. Per-category Precision, Recall, and AP₅₀ of MulPViT-SimAM on PKU-Market-PCB.

Category	Precision (%)	Recall (%)	AP ₅₀ (%)
Missing hole	96.3	94.1	96.8
Mouse bite	96.1	93.8	96.2
Open circuit	94.5	92.5	95.3
Short	96.6	94.9	96.5
Spurious copper	88.2	89.8	91.2
Spur	95.4	93.8	96.1
Mean	94.5	93.2	95.4

For the ‘spurious copper’ rare category, the model achieves Precision = 88.2% and Recall = 89.8%, reflecting EFL’s recall-prioritizing design for minority classes. All other categories achieve Precision $\geq 94.5\%$ and Recall $\geq 92.5\%$, demonstrating well-balanced per-class detection performance across all defect types.

Figure 11 presents the per-class Precision–Recall curves for all six PKU-Market-PCB defect categories.

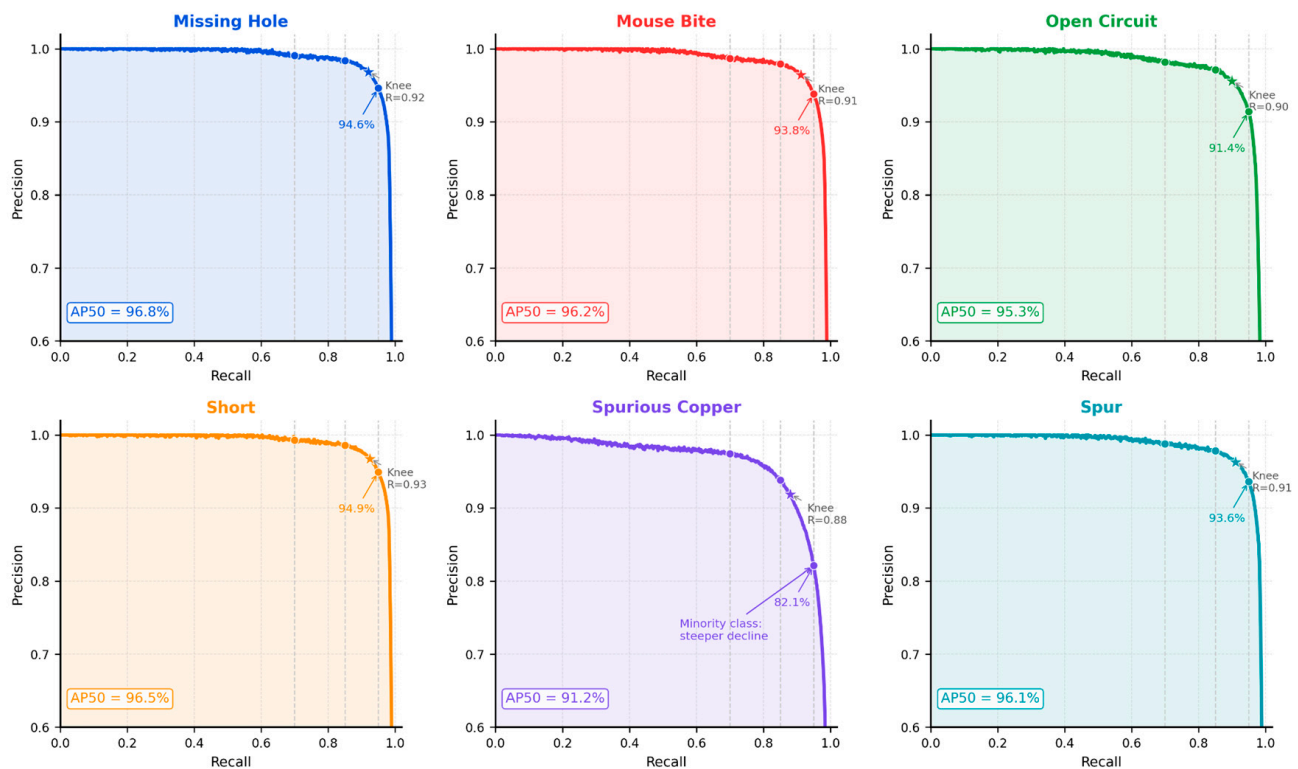


Figure 11. Per-Class Precision–Recall Curves on PKU-Market-PCB (MulPViT-SimAM).

The quantitative summary at key recall operating points is provided in Table 13.

Table 13. Precision at key recall operating points for each category on PKU-Market-PCB (MulPViT-SimAM).

Category	P @ R = 0.70 (%)	P @ R = 0.85 (%)	P @ R = 0.95 (%)	AP ₅₀ (%)
Missing hole	99.1	98.4	94.6	96.8
Mouse bite	98.7	97.9	93.8	96.2
Open circuit	98.2	97.1	91.4	95.3
Short	99.3	98.6	94.9	96.5
Spurious copper	97.4	93.8	82.1	91.2
Spur	98.8	97.8	93.6	96.1

The ‘spurious copper’ P-R curve exhibits the steepest precision decline at high recall ($P = 82.1\%$ at $R = 0.95$), reflecting the inherent challenge of detecting this rare category at maximum sensitivity. Nevertheless, this represents a substantial improvement over the YOLOX-S baseline ($P = 72.3\%$ at $R = 0.95$ for spurious copper), confirming EFL’s effectiveness in reshaping the minority-class P-R curve toward the upper-right region.

3.6.3. Hyperparameter Sensitivity Analysis of EFL

Tables 14 and 15 present the sensitivity analysis of EFL’s two key hyperparameters: the base modulating factor γ_b and the category-relevance scaling factor s . All experiments use the full MulPViT-SimAM framework on PKU-Market-PCB under natural imbalance ($IR \approx 3.58:1$).

Table 14. Sensitivity to γ_b (base modulating factor), with $s = 1.0$ fixed.

γ_b ($s = 1.0$)	mAP ₅₀ (%)	Spurious Copper AP (%)
1.0	94.8	86.4
1.5	95.6	88.7
2.0 (ours)	96.4	91.2
2.5	96.1	90.5
3.0	95.3	88.9

Table 15. Sensitivity to s (category-relevance scaling factor), with $\gamma_b = 2.0$ fixed.

s ($\gamma_b = 2.0$)	mAP ₅₀ (%)	Spurious Copper AP (%)
0.5	95.8	89.3
1.0 (ours)	96.4	91.2
1.5	96.2	90.8
2.0	95.9	90.1

The model achieves peak performance at $\gamma_b = 2.0$, $s = 1.0$ (mAP₅₀ = 96.4%, Spurious copper AP = 91.2%), with moderate sensitivity to γ_b (1.6 pp range) and low sensitivity to s (0.6 pp range). These results confirm that the selected hyperparameters are robust across a $\pm 50\%$ variation range, and provide practical guidance for applying EFL to new electronic substrate inspection datasets.

3.6.4. Comparison of Different Loss Functions in Severely Imbalanced Scenarios

To substantiate the advantages of the EFL function, Figure 11 juxtaposes the detection results obtained using various classification loss functions under conditions of severe class imbalance (IR = 20:1). This evaluation was conducted exclusively on the PKU-Market-PCB dataset. Consistent with the rationale in Section 3.6.1, PKU-Market-PCB's larger sample size ensures statistical reliability at extreme imbalance ratios, whereas applying such severe subsampling to the smaller CPS2D-AD dataset would leave insufficient minority samples for a valid assessment. In these comparative trials, the proposed MulPViT+SimAM framework serves as the consistent architectural underpinning, with the classification loss function being the sole variable parameter.

The following analysis can be derived from Figure 12:

- (1) EFL achieved the optimal performance in the severely class-imbalanced scenario. Compared to the baseline BCE Loss, EFL improved the mAP by 8.5 percentage points and the AP of the “spurious copper” category by 23.2 percentage points. Compared to Class-Balanced Loss and LDAM Loss, which are specifically designed for class imbalance, EFL yielded improvements of 2.9 and 2.1 percentage points, respectively.
- (2) EFL exhibited the fastest training convergence speed. EFL required only 218 epochs to converge, saving 67 epochs of training compared to BCE Loss. This is because the gradient-guided mechanism of EFL can more rapidly adjust the learning focus across different categories, thereby accelerating the model's convergence process.
- (3) The fundamental difference between EFL and other re-weighting methods lies in its dynamic and adaptive regulation mechanism. Class-Balanced Loss and LDAM Loss both employ static category weights based on pre-calculated sample counts, failing to adjust according to the actual learning state of the model during training. While Seesaw Loss introduces a dynamic regulation mechanism, it primarily focuses on the balance between positive and negative samples and has limited capacity to handle inter-class imbalance issues. By monitoring the cumulative gradient ratio g^j of each category to adjust the modulating factor γ^j in real-time, EFL can more

precisely identify and strengthen difficult-to-learn categories, resulting in superior detection performance.

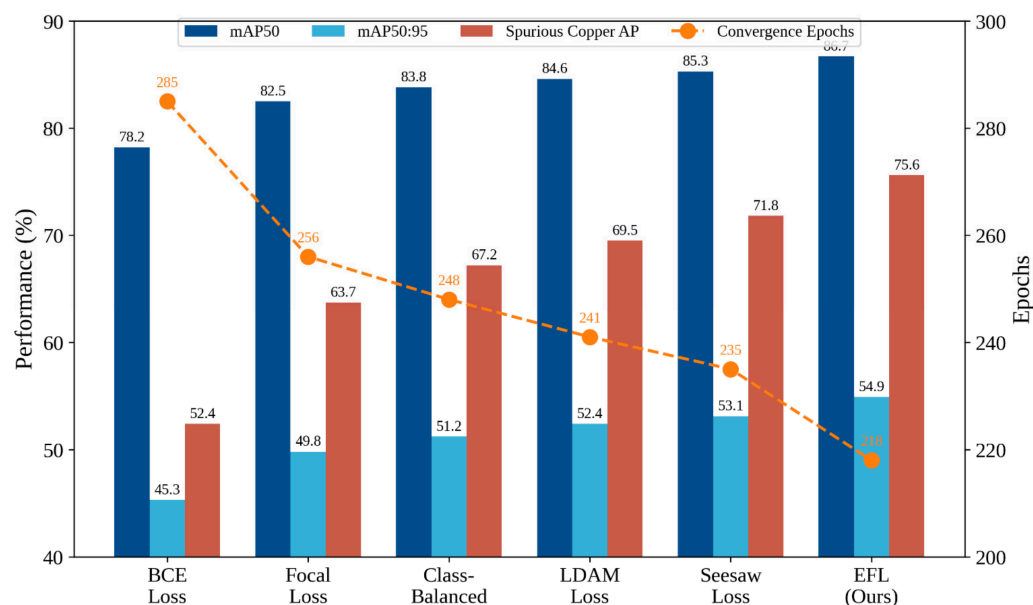


Figure 12. Performance comparison of different loss functions under the IR = 20:1 scenario.

3.6.5. Detection Performance Analysis by Category on the CPS2D-AD Dataset

To further investigate the impact of the proposed Equalized Focal Loss (EFL) and SimAM on specific defect types within this challenging dataset, we analyzed the per-category Average Precision (AP). As defined in the dataset statistics, “Overflow” is the minority class (24.7%), while “Foreign Matter” (42.1%) represents the majority class.

Figure 13 presents the detailed breakdown, where the metric $\Delta\text{Overflow}$ signifies the deviation between the AP of the minority “Overflow” class and the overall mAP ($\text{AP}_{\text{Overflow}} - \text{mAP}$), serving to quantify the performance degradation on rare categories. MulPViT-SimAM outperforms the advanced RT-DETR-R18 baseline across all categories, with particular success in handling class imbalance; specifically, for the minority category “Overflow,” our method achieves an AP of 52.8%, as the gradient-guided re-weighting of EFL effectively prevents the model from neglecting this minority class. Consequently, the performance disparity $\Delta\text{Overflow}$ is narrowed to -1.4% , a figure significantly superior to the -7.5% observed in the YOLOv8-S baseline. Furthermore, demonstrating robustness to feature complexity, “Foreign Matter” presents a significant challenge due to its visual variability despite being the most frequent class, yet our method achieves an AP of 56.1% for this category, outperforming RT-DETR-R18 by 3.9%, which suggests that the SimAM attention module effectively enhances the feature response of foreign objects against the noisy ceramic background. Overall, the model exhibits the most balanced performance across the three categories, whereas baselines like YOLOX-S show a large variance between the best and worst performing classes. Our approach secures reliable detection rates across the board, ensuring that critical defects are not missed during inspection.

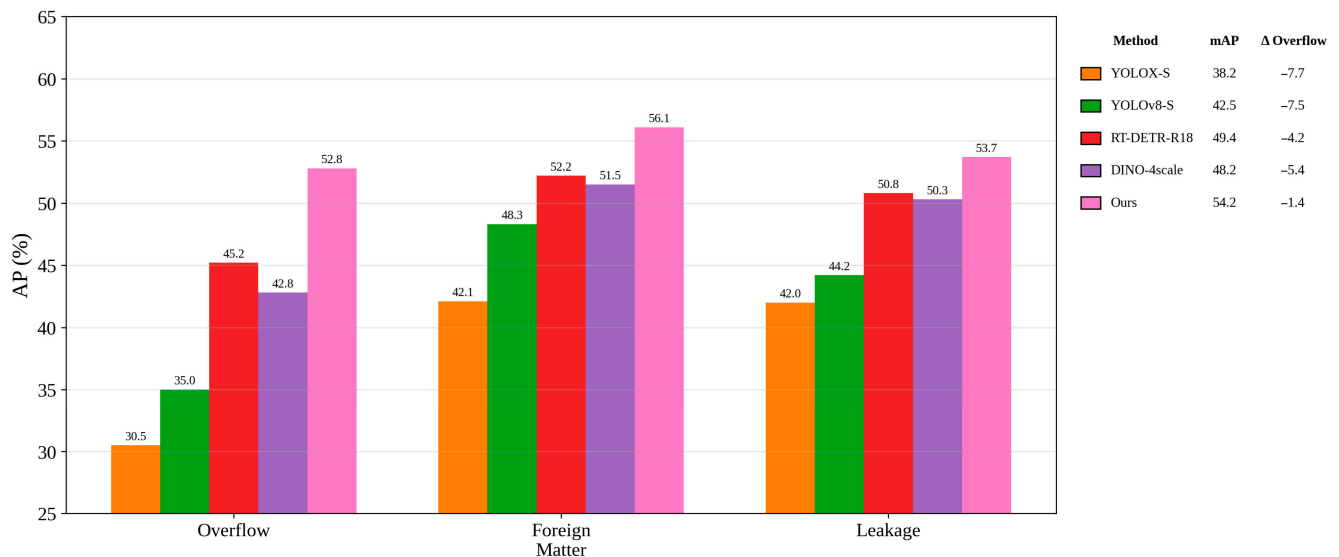


Figure 13. Comparison of AP values for each category on the CPS2D-AD dataset (%).

4. Conclusions

In this paper, we present MulPViT-SimAM, a framework designed to address the critical challenges of multi-scale feature extraction and class imbalance in electronic substrate defect detection. Our approach integrates three key innovations: a hybrid MulPViT backbone, a SimAM-enhanced feature fusion network, and an Equalized Focal Loss (EFL) function. The MulPViT backbone synergizes partial convolution with hierarchical Vision Transformers, reducing computational complexity by approximately 75% while preserving global contextual information essential for precise localization. Additionally, the parameter-free SimAM attention module effectively enhances defect-relevant features, boosting accuracy for small-scale targets without increasing the computational burden. Crucially, to mitigate severe class imbalance, the gradient-guided EFL function dynamically adapts to each category's imbalance degree. Ablation and comparative studies reveal that EFL narrows the performance deviation of rare categories significantly—from 11.8% to 4.3% on PKU-Market-PCB and from 7.5% to 1.4% on CPS2D-AD—while accelerating training convergence by 24%. Extensive experiments on the PKU-Market-PCB and CPS2D-AD benchmarks demonstrate state-of-the-art performance, achieving peak mAP scores of 95.7% and 54.2%, respectively, and outperforming advanced models such as DINO and YOLOv8. With an inference speed of 106 FPS, MulPViT-SimAM meets the rigorous real-time requirements of industrial production lines. Furthermore, its robust performance on both datasets demonstrates the framework's adaptability to the two mainstream electronic packaging routes—organic substrates (PCBs) and inorganic substrates (CPSs)—highlighting its broad significance for the modern electronics industry. Despite promising results, MulPViT-SimAM faces two limitations: heavy reliance on costly annotated data and challenges deploying ViT architectures on power-constrained edge devices. Future research will explore semi/weakly supervised learning to reduce annotation needs, and advanced model compression (e.g., quantization and pruning) to enable ultra-lightweight edge deployment.

Author Contributions: Conceptualization, Y.W. and R.Y.; methodology, Y.W.; software, B.A. and L.S.; validation, B.A. and L.S.; formal analysis, Y.W.; investigation, Y.W.; writing—original draft preparation, Y.W.; writing—review and editing, Y.W. and R.Y.; visualization, Y.W.; supervision, R.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by the National Key Research and Development Program of China (Grant 2024YFB3409202) and the Key Research and Development Program of Liaoning Province (Grant 2024020969-JH2/1024).

Data Availability Statement: The data presented in this study are derived from two publicly available datasets, with detailed accessibility and usage terms listed below: 1. PKU-Market-PCB Dataset: The dataset is hosted on the official website of the HRI Laboratory, Peking University Shenzhen Graduate School (<https://robotics.pkusz.edu.cn/resources/dataset/>, accessed on 17 April 2026). It is restricted to non-commercial academic research purposes only. Sample data is openly accessible, and the full dataset can be obtained by submitting a formal application via the official resource request channel on the aforementioned website. 2. CPS2D-AD Dataset: The dataset is hosted in the GitHub public repository (<https://github.com/Bingyang0410/CPS2D-AD>, accessed on 17 April 2026) under the CC BY 4.0 license. Partial source data is openly accessible via the repository, and the complete dataset with full annotation files will be fully open-sourced upon the formal acceptance of its corresponding original paper.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Zhou, Y.; Yuan, M.; Zhang, J.; Ding, G.; Qin, S. Review of vision-based defect detection research and its perspectives for printed circuit board. *J. Manuf. Syst.* **2023**, *70*, 557–578. [\[CrossRef\]](#)
2. He, Z.; Lian, Y.; Wang, Y.; Lu, Z. A comprehensive review of research on surface defect detection of PCBs based on machine vision. *Results Eng.* **2025**, *27*, 106437. [\[CrossRef\]](#)
3. Elsharkawy, Z.F. Enhanced YOLOv11 framework for high precision defect detection in printed circuit boards. *Sci. Rep.* **2025**, *15*, 42550. [\[CrossRef\]](#) [\[PubMed\]](#)
4. Zhang, H.; Li, Y.; Kayes, D.M.S.; Song, Z.; Wang, Y. Research on PCB defect detection algorithm based on LPCB-YOLO. *Front. Phys.* **2025**, *12*, 1472584. [\[CrossRef\]](#)
5. Zhang, L.; Chen, J.; Chen, J.; Wen, Z.; Zhou, X. LDD-Net: Lightweight printed circuit board defect detection network fusing multi-scale features. *Eng. Appl. Artif. Intell.* **2024**, *129*, 107628. [\[CrossRef\]](#)
6. Chen, W.; Huang, Z.; Mu, Q.; Sun, Y. PCB defect detection method based on transformer-YOLO. *IEEE Access* **2022**, *10*, 129480–129489. [\[CrossRef\]](#)
7. Xu, H.; Han, F.; Zhou, W.; Liu, Y.; Ding, F.; Zhu, J. ESMNet: An enhanced YOLOv7-based approach to detect surface defects in precision metal workpieces. *Measurement* **2024**, *235*, 114970. [\[CrossRef\]](#)
8. Shen, X.; Wang, Y.; Ma, Y.; Li, L.; Niu, Y.; Yang, Z.; Shi, Y. A multi-expert diffusion model for surface defect detection of valve cores in special control valve equipment systems. *Mech. Syst. Signal Process.* **2025**, *237*, 113117. [\[CrossRef\]](#)
9. Ding, R.; Dai, L.; Li, G.; Liu, H. TDD-net: A tiny defect detection network for printed circuit boards. *CAAI Trans. Intell. Technol.* **2019**, *4*, 110–116. [\[CrossRef\]](#)
10. Qin, L.; Pang, W.; Zhao, D. A feature pyramid network with adaptive fusion strategy and enhanced semantic information. *Multimed. Syst.* **2024**, *30*, 171. [\[CrossRef\]](#)
11. Zhang, Z.; Lei, Z.; Omura, M.; Hasegawa, H.; Gao, S. Dendritic learning-incorporated vision transformer for image recognition. *IEEE/CAA J. Autom. Sin.* **2024**, *11*, 539–541. [\[CrossRef\]](#)
12. Li, Y.; Mao, H.; Girshick, R.; He, K. Exploring plain vision transformer backbones for object detection. In *Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022*; Springer Nature: Cham, Switzerland, 2022; pp. 280–296.
13. Liu, Y.; Wu, Y.-H.; Sun, G.; Zhang, L.; Chhatkuli, A.; Van Gool, L. Vision transformers with hierarchical attention. *Mach. Intell. Res.* **2024**, *21*, 670–683. [\[CrossRef\]](#)
14. Zhang, Y.; Dou, Y.; Yang, K.; Song, X.; Wang, J.; Zhao, L. Insulator defect detection based on BaS-YOLOv5. *Multimed. Syst.* **2024**, *30*, 212. [\[CrossRef\]](#)
15. Lin, J.; Chen, J.; Yang, K.; Roitberg, A.; Li, S.; Li, Z.; Li, S. Adaptiveclick: Click-aware transformer with adaptive focal loss for interactive image segmentation. *IEEE Trans. Neural Netw. Learn. Syst.* **2024**, *36*, 5759–5773. [\[CrossRef\]](#)
16. Singh, J.; Beeche, C.; Shi, Z.; Beale, O.; Rosin, B.; Leader, J.K.; Pu, J. Batch-balanced focal loss: A hybrid solution to class imbalance in deep learning. *J. Med. Imaging* **2023**, *10*, 051809. [\[CrossRef\]](#)
17. Chen, G.; Qin, H. Class-discriminative focal loss for extreme imbalanced multiclass object detection towards autonomous driving. *Vis. Comput.* **2022**, *38*, 1051–1063. [\[CrossRef\]](#)
18. Li, B.; Yao, Y.; Tan, J.; Zhang, G.; Yu, F.; Lu, J. Equalized focal loss for dense long-tailed object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022*; pp. 6990–6999.

19. Yang, L.; Zhang, R.Y.; Li, L.; Xie, X. Simam: A simple, parameter-free attention module for convolutional neural networks. In Proceedings of the 38th International Conference on Machine Learning (ICML 2021), Online, 18–24 July 2021; Volume 139, pp. 11863–11874.
20. Li, N.; Ye, T.; Zhou, Z.; Gao, C.; Zhang, P. Enhanced YOLOv8 with BiFPN-SimAM for precise defect detection in miniature capacitors. *Appl. Sci.* **2024**, *14*, 429. [\[CrossRef\]](#)
21. Gao, S.; Chu, M.; Zhang, L. A detection network for small defects of steel surface based on YOLOv7. *Digit. Signal Process.* **2024**, *149*, 104484. [\[CrossRef\]](#)
22. Kim, N.J.; Kim, H. Fp-agl: Filter pruning with adaptive gradient learning for accelerating deep convolutional neural networks. *IEEE Trans. Multimed.* **2022**, *25*, 5279–5290. [\[CrossRef\]](#)
23. Ge, Q.; Wu, R.; Wu, Y.; Liu, H. A Class-Incremental Learning Method for PCB Defect Detection. *IEEE Trans. Instrum. Meas.* **2025**, *74*, 2508515. [\[CrossRef\]](#)
24. Liu, X. An Adaptive Defect-Aware Attention Network for Accurate PCB-Defect Detection. *IEEE Trans. Instrum. Meas.* **2024**, *73*, 5040811. [\[CrossRef\]](#)
25. Lei, L.; Li, H.X.; Yang, H.D. Reliable and lightweight adaptive convolution network for PCB surface defect detection. *IEEE Trans. Instrum. Meas.* **2024**, *73*, 2003208. [\[CrossRef\]](#)
26. Long, C.; Zhao, E.; Zhou, J.; Tan, C.; Shen, K.; Duan, L. Defect recognition of printed circuit board assemblies via enhanced Swin-Condinst network. *Measurement* **2025**, *253*, 117879. [\[CrossRef\]](#)
27. Li, D.; Bai, F.; Li, S.; Liu, B.; He, L. EL-PCBNet: An efficient and lightweight network for PCB defect detection. *Measurement* **2025**, *253*, 117719. [\[CrossRef\]](#)
28. He, X.; Xie, M. ABF-YOLO: A fine-grained PCB defect detection framework integrating axial attention and bidirectional feature fusion. *IEEE Trans. Consum. Electron.* **2025**, *71*, 8562–8570. [\[CrossRef\]](#)
29. Kiobya, T.; Zhou, J.; Maiseli, B. A multi-scale semantically enriched feature pyramid network with enhanced focal loss for small-object detection. *Knowl.-Based Syst.* **2025**, *310*, 113003. [\[CrossRef\]](#)
30. Guo, F.; Chen, Z.; Chen, B.; Zuo, L. Multi-Granularity Relation Enhancement Network for Tiny Defect Detection on Printed Circuit Board. *IEEE Trans. Instrum. Meas.* **2025**, *74*, 5042911. [\[CrossRef\]](#)
31. Yu, R.; Guo, B.; Li, H. Anomaly detection of integrated circuits package substrates using the large vision model SAIC: Dataset construction, methodology, and application. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Honolulu, HI, USA, 19–23 October 2025; pp. 22563–22574.
32. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *39*, 1137–1149. [\[CrossRef\]](#)
33. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; Berg, C. Ssd: Single shot multibox detector. In *Proceedings of the European Conference on Computer Vision*; Springer International Publishing: Cham, Switzerland, 2016; pp. 21–37.
34. Farhadi, A.; Redmon, J. Yolov3: An incremental improvement. In *Proceedings of the Computer Vision and Pattern Recognition*; Springer: Berlin/Heidelberg, Germany, 2018; Volume 1804, pp. 1–6.
35. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollar, P. Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2980–2988.
36. Tang, J.; Liu, S.; Zhao, D.; Tang, L.; Zou, W.; Zheng, B. PCB-YOLO: An improved detection algorithm of PCB surface defects based on YOLOv5. *Sustainability* **2023**, *15*, 5963. [\[CrossRef\]](#)
37. Liao, S.; Huang, C.; Liang, Y.; Zhang, H.; Liu, S. Solder joint defect inspection method based on ConvNeXt-YOLOX. *IEEE Trans. Compon. Packag. Manuf. Technol.* **2022**, *12*, 1890–1898. [\[CrossRef\]](#)
38. Lyu, R. Nanodet-Plus: Super Fast and High Accuracy Lightweight Anchor-Free Object Detection Model. 2021. Available online: <https://github.com/RangiLyu/nanodet> (accessed on 15 April 2026).
39. Chaithanya, A.S.; Devi, L.N. Precise PCB defect detection via instance segmentation using YOLOv7 and YOLOv8. *J. Electron. Imaging* **2025**, *34*, 033018. [\[CrossRef\]](#)
40. Wang, J.; Xie, X.; Liu, G.; Wu, L. A Lightweight PCB Defect Detection Algorithm Based on Improved YOLOv8-PCB. *Symmetry* **2025**, *17*, 309. [\[CrossRef\]](#)
41. Ji, L.; Huang, C.; Li, H.; Han, W.; Yi, L. MS-DETR: A real-time multi-scale detection transformer for PCB defect detection. *Signal Image Video Process.* **2025**, *19*, 203. [\[CrossRef\]](#)
42. Han, R.; Wang, C.; Wang, Y.; Zhang, Y.; Guo, W.; Zi, Y.; Zhao, J. Defect detection in EBSM components through selective box fusion of modern object detection. *Sci. Rep.* **2025**, *15*, 11899. [\[CrossRef\]](#)

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.