American International Journal of Scientific & Technical Research

Official Website: https://aijstr.com

Lightweight Neural Network Optimization for Real-Time Industrial Image Detection

Brandon J. Miller, Ethan S. Walker, Lucas Reed Phillips III

A B S T R A C T

Intelligent industrial inspection based on machine vision technology has become an indispensable core technology in modern intelligent manufacturing and industrial digital transformation. With the rapid popularization of industrial edge computing equipment and automated production lines, industrial quality inspection has put forward higher comprehensive requirements for detection accuracy, real-time inference performance, model deployment adaptability, and environmental robustness (Lee & Park, 2022). Traditional manual inspection methods are limited by low efficiency, strong subjective judgment, fatigue error and high labor cost, which cannot meet the high-frequency and high-precision detection needs of modern large-scale industrial production. Traditional convolutional neural network-based visual detection models have outstanding advantages in feature learning and defect recognition, but they generally have prominent defects such as excessive parameter redundancy, huge floating-point computation, high inference delay, and poor adaptability to low-power embedded edge devices (Sandler et al., 2018). In actual industrial deployment scenarios, most high-precision deep learning models rely on high-performance GPU computing resources, which are costly, high-power consumption, and difficult to deploy on a large scale in industrial field edge terminals with limited hardware resources.

To solve the inherent contradiction between detection accuracy and real-time performance in industrial multi-class surface defect detection tasks, and break through the deployment bottleneck of traditional neural networks in resource-constrained industrial edge scenarios, this study proposes an improved lightweight industrial detection network (IL-IDNet) tailored for complex industrial real-time image detection tasks. Aiming at the typical characteristics of industrial defects such as sparse distribution, tiny target scale, low contrast, irregular morphology, and easy feature flooding by background noise, the proposed model carries out targeted structural optimization and mechanism innovation. Firstly, an industrial adaptive improved depthwise separable convolution module is designed to replace the traditional generic convolution unit, which dynamically enhances the feature response of defect targets and suppresses invalid background texture interference, effectively improving the model's sensitivity to micro and low-contrast industrial defects. Secondly, a structured network pruning strategy based on feature contribution degree is adopted to completely remove redundant network layers and invalid parameter calculation, which realizes model lightweight compression without damaging core feature extraction capability. Thirdly, a streamlined bidirectional multi-scale feature fusion mechanism is constructed to screen high-value cross-scale feature information, eliminate redundant fusion links, and realize efficient complementary fusion of shallow fine-grained micro-defect features and deep high-semantic large-scale defect features.

In order to verify the comprehensive performance and engineering practicability of the proposed IL-IDNet model, this paper conducts sufficient quantitative and qualitative comparison experiments based on the public authoritative NEU-DET industrial defect benchmark dataset and the self-collected real industrial complex scene dataset. Multiple classic mainstream lightweight networks including MobileNetV3, ShuffleNetV2 and SqueezeNet are selected as baseline models for horizontal comparison from multiple dimensions such as detection accuracy, model parameter scale, computational complexity, inference speed, and environmental anti-interference ability. The experimental results show that the proposed IL-IDNet model achieves a mean average precision (mAP@0.5) of 95.21% in multi-class industrial defect detection, which is significantly higher than all comparison baseline models. Meanwhile, the model parameter size is only 2.8M, and the floating-point operation overhead is far lower than that of traditional lightweight networks. In the actual test of low-power industrial embedded edge terminals, the average inference frame rate reaches 68.5 FPS, which far exceeds the 30 FPS real-time detection standard required by industrial online production lines. In complex industrial interference environments such as uneven illumination, cluttered background texture and strong noise, the proposed model maintains extremely low missed detection rate and false detection rate for tiny defects and low-contrast defects, showing excellent environmental robustness and cross-scene generalization ability.The comprehensive experimental verification fully proves that the improved lightweight network proposed in this study realizes the optimal balanced optimization of detection accuracy, lightweight degree and real-time inference speed, effectively solves the performance imbalance problem of traditional lightweight models in industrial detection tasks. The optimized model has strong deployment adaptability and engineering practical value, which can be widely applied to edge intelligent inspection, real-time production quality supervision, automated defect screening and other scenarios in intelligent manufacturing workshops, and provides a reliable technical reference for the intelligent upgrading and

Corresponding author : Brandon J. Miller

efficient operation of modern industrial production systems.

Keywords: *Industrial Defect Detection; Lightweight Neural Network; Edge Computing; Feature Fusion; Structured Pruning; Machine Vision; Real-Time Detection*

1. Introduction

With the in-depth advancement of Industry 4.0 strategy and industrial intelligent transformation worldwide, automated and intelligent detection technology has completely replaced traditional manual inspection and become the core guarantee of industrial product quality control and production safety supervision (Lu & Lee, 2023). Modern industrial production presents the characteristics of high speed, batch scale, diversification and high precision. A large number of mechanical parts, metal surfaces, electronic components and industrial finished products will inevitably produce various surface defects such as micro-scratches, corrosion pits, cracks, deformation and stains in the process of production, processing and assembly. These tiny defects will gradually expand with the service process of products, leading to product performance degradation, service life reduction, and even serious industrial safety accidents in severe cases (Lebedev & Lempitsky, 2016). Therefore, efficient, accurate and real-time industrial surface defect detection is a key link to ensure product qualification rate, stabilize production quality and reduce industrial operation risks.

Traditional industrial defect detection methods are mainly divided into manual visual inspection and traditional machine vision algorithm detection. Manual inspection relies on the subjective experience of inspectors, which has obvious limitations such as low detection efficiency, high labor cost, easy visual fatigue, strong subjectivity and inconsistent detection standards. It is impossible to meet the ultra-high frequency and large-scale detection requirements of modern automated production lines (Alfatemi et al., 2026). Traditional machine vision detection algorithms are based on artificial feature design and fixed threshold segmentation, including edge detection, contour extraction, threshold binarization and other technical means. Such algorithms can only achieve stable detection for single scene, single defect type and fixed lighting environment, and have extremely poor generalization ability for complex industrial scenes with variable illumination, complex background and diverse defect morphologies. Once the production environment or product parameters change, the algorithm needs to be manually adjusted and optimized repeatedly, which has high maintenance cost and low intelligence, and is difficult to adapt to the flexible production mode of modern industry.

In recent years, the rapid development of deep learning technology has brought revolutionary breakthroughs to the field of industrial visual detection. Convolutional Neural Network (CNN) has powerful autonomous feature learning, high-dimensional semantic representation and multi-target feature recognition capabilities, which can automatically mine hidden deep feature information of defects from massive industrial image data, and overcome the defects of poor generalization and low intelligence of traditional manual feature design algorithms (Howard et al., 2017). At present, a series of high-precision deep learning detection models such as Faster R-CNN, YOLO series and EfficientNet have been widely used in industrial defect detection research, and have achieved excellent detection accuracy in laboratory simulation environments. However, most high-precision models have complex network structures, huge parameter scales and massive floating-point computation, which rely on high-performance GPU hardware support and cannot be deployed on low-cost, low-power industrial embedded edge devices widely used in industrial sites (Gholami et al., 2021). Industrial edge terminals such as industrial control boards, embedded detection modules and mobile monitoring equipment are limited by hardware conditions such as small memory, low computing power and insufficient power supply, which cannot bear the high computational overhead of large-scale deep learning models. This forms a serious deployment contradiction between high-precision detection requirements and limited edge hardware resources, and restricts the large-scale industrial application of deep learning visual detection technology.

Lightweight neural network optimization technology is the core way to solve the edge deployment problem of deep learning models in industrial scenarios. The mainstream lightweight network architectures represented by MobileNet, ShuffleNet and SqueezeNet reduce model parameters and computational complexity through depthwise separable convolution, channel shuffling and fire module compression mechanisms, which greatly reduce the hardware resource consumption of models (Iandola et al., 2016; Zhang et al., 2018). However, the existing classic lightweight networks are all optimized for general natural image datasets such as ImageNet, and there are serious mismatches between network structure design and industrial defect detection task characteristics. Industrial defects have unique characteristics different from natural image targets: the defect area is small in proportion, the foreground and background contrast is low, the defect morphology is irregular, the distribution is sparse, and it is easy to be submerged by complex background noise and light interference (Su et al., 2026). When generic lightweight networks are directly applied to industrial defect detection, they are prone to problems such as insufficient micro-defect feature extraction capability, serious feature loss, and high misdetection and missed detection rates.

In addition, the current industrial lightweight network optimization research generally has two extreme technical defects. On the one hand, some studies simply adopt massive network pruning and parameter compression to pursue ultra-lightweight model volume, which directly leads to the attenuation of model feature extraction capability, especially the serious loss of tiny defect detail features, resulting in a sharp decline in detection accuracy (Mao et al., 2018). On the other hand, some studies add complex attention modules, multi-scale fusion branches and deep network structures to improve detection accuracy, which greatly increases model computational complexity and inference delay, cannot meet the real-time detection requirements of industrial high-speed production lines (Bhuiyan et al., 2025). At present, there is a lack of a systematic lightweight optimization scheme that can balance model lightweight degree, detection accuracy and real-time inference speed for complex industrial multi-scale defect detection scenarios, which has become an important technical bottleneck restricting the engineering popularization of industrial intelligent detection systems.

In view of the above research gaps and technical deficiencies, this paper proposes an improved lightweight industrial real-time detection network (IL-IDNet) and forms a complete set of targeted optimization mechanisms. The core research objectives of this paper are as follows: first, to solve the problem of poor feature extraction ability of traditional lightweight networks for low-contrast micro-defects in complex industrial environments; second, to eliminate the redundant computation and parameter waste of existing lightweight models in industrial detection tasks; third, to realize the organic balance of model lightweight degree, detection precision and inference real-time performance; fourth, to build a lightweight detection model with strong industrial scene adaptability and edge deployment performance. The research results of this paper can effectively promote the landing application of deep learning visual detection technology in resource-constrained industrial edge scenarios, and provide new ideas and technical support for the intelligent and efficient upgrading of industrial quality inspection links.

2. Related Work

2.1 Development and Limitations of Traditional Industrial Visual Detection Algorithms

Industrial visual detection technology has experienced three important development stages: manual detection, traditional machine vision algorithm detection and deep learning intelligent detection (Leu et al., 2015). In the early stage of industrial development, product quality inspection completely relied on manual visual observation and manual screening, which had extremely low detection efficiency and could only meet the production and inspection needs of small-batch and low-precision industrial products. With the continuous improvement of industrial production automation, traditional machine vision algorithms based on artificial feature design have gradually replaced manual detection and become the mainstream industrial detection technology for a long time. Common traditional detection algorithms include Canny edge detection, Sobel gradient detection, Otsu threshold segmentation, template matching and other technical types (Xu et al., 2019). These algorithms have simple principles and fast detection speed for fixed single scenes, and are widely used in simple industrial defect detection tasks such as surface scratch detection and foreign object identification.

However, with the diversification of industrial product types, the complexity of production environment and the improvement of defect detection precision requirements, the inherent limitations of traditional machine vision algorithms are increasingly prominent. All traditional algorithms rely on manual design of feature rules and artificial adjustment of threshold parameters, which require professional technicians to carry out targeted algorithm debugging for different products and different defect types. The algorithm has no autonomous learning ability and cannot automatically adapt to the changes of industrial lighting, background texture and defect scale. In complex industrial scenarios with uneven illumination, strong background noise and irregular defect morphology, traditional algorithms are easy to produce feature misjudgment and target missing, with poor detection stability and low accuracy (Sandler et al., 2018). In addition, traditional machine vision algorithms can only achieve detection for single defect type, and cannot realize simultaneous identification and classification of multi-scale and multi-type industrial defects, which is difficult to meet the comprehensive quality supervision needs of modern integrated industrial production.

2.2 Research Progress of Deep Learning Detection Models

Deep learning-based visual detection algorithms have achieved disruptive breakthroughs in target detection and defect recognition tasks, and gradually become the mainstream research direction of industrial intelligent detection. According to the network structure and detection process, deep learning detection models are divided into two-stage detection frameworks and one-stage detection frameworks (Redmon & Farhadi, 2018). The two-stage detection framework represented by Faster R-CNN completes target detection through two steps of candidate region generation and feature classification. It has the advantages of high detection accuracy and strong small-target recognition ability, and can achieve stable detection for tiny industrial defects (Ren et al., 2015). However, the two-stage detection mechanism leads to complex model reasoning process, large computational overhead and slow inference speed, which cannot meet the real-time detection requirements of industrial high-speed assembly lines and online production monitoring.

The one-stage detection framework represented by YOLO and SSD integrates feature extraction, target positioning and category classification into a single forward propagation process, which omits the independent candidate region generation link of the two-stage model, greatly simplifies the model reasoning process, and significantly improves the real-time detection performance (Bhuiyan et al., 2025). The one-stage detection model has fast inference speed and simple deployment process, which is more suitable for industrial online real-time detection scenarios. Nevertheless, the original YOLO and SSD series models have large parameter scales, complex network structures and high hardware resource requirements. Although the real-time performance is improved compared with the two-stage model, they still cannot be deployed on ultra-low-power industrial embedded edge devices, and there are still serious deployment limitations in resource-constrained industrial scenarios.

2.3 Research Status and Deficiencies of Classic Lightweight Neural Networks

In order to solve the deployment problem of deep learning models on mobile terminals and edge devices, international scholars have proposed a series of classic lightweight neural network optimization models in recent years, which mainly realize model lightweight through convolution structure improvement, channel compression and parameter pruning (Howard et al., 2017). SqueezeNet proposed by Iandola et al. (2016) innovatively designs squeeze-and-expand fire modules, which compress and expand feature channels to reduce model parameters. The model size is only 0.5MB, which achieves detection accuracy equivalent to AlexNet with extremely low parameter volume, and greatly reduces the threshold of model hardware deployment. MobileNet series networks optimize the standard convolution structure and propose depthwise separable convolution, which splits the standard convolution into depthwise convolution and pointwise convolution, effectively reducing floating-point operations and parameter redundancy (Sandler et al., 2018). MobileNetV2 adds inverted residual structure and linear bottleneck layer on the basis of the original network, which further improves the feature representation ability of lightweight models. MobileNetV3 introduces lightweight attention mechanism and adaptive activation function, which realizes a better balance between model accuracy and speed.

ShuffleNet series networks proposed by Zhang et al. (2018) solve the problem of information isolation in group convolution through channel shuffling operation, realize efficient information interaction between different convolution groups, and maintain feature extraction capability under ultra-low parameter and computation conditions. The above classic lightweight networks have excellent performance in general natural image classification and daily target detection tasks, but they have obvious inadaptability when applied to industrial defect detection scenarios (Alfatemi et al., 2026). On the one hand, the feature extraction design of generic lightweight networks is oriented to large-scale obvious targets in natural images, and the feature capture ability for tiny, low-contrast industrial defects is insufficient. On the other hand, the network anti-interference design is not aimed at industrial complex noise and variable illumination environment, and it is easy to produce feature attenuation and detection failure in industrial scenes. In addition, the multi-scale feature fusion mechanism of traditional lightweight networks is single, which cannot adapt to the multi-scale distribution characteristics of industrial defects, resulting in unbalanced detection performance for large-area defects and micro-defects.

2.4 Research Progress and Existing Problems of Industrial-Oriented Lightweight Optimization

In view of the application limitations of generic lightweight networks in industrial scenarios, many scholars at home and abroad have carried out targeted research on industrial lightweight network optimization in recent years (Gholami et al., 2021). Some studies optimize the network quantization strategy, compress the model storage space through low-bit quantization, and reduce the computational overhead of model edge deployment. Some studies adopt group pruning and sparse optimization methods to remove redundant model parameters and improve model inference efficiency. Some studies add lightweight attention modules to the network to enhance the feature response of defect targets and improve detection accuracy (Lebedev & Lempitsky, 2016). These research results have effectively promoted the development of industrial lightweight detection technology, but there are still many unresolved technical problems.

Through comprehensive literature analysis and experimental verification, the existing industrial lightweight optimization research has three core deficiencies. First, most optimization schemes adopt single lightweight or single accuracy improvement strategy, which cannot balance the lightweight degree and detection performance of the model. Simple compression leads to accuracy loss, and simple accuracy improvement leads to reduced real-time performance (Su et al., 2026). Second, most industrial lightweight models lack targeted structural design for industrial defect multi-scale characteristics, and have poor adaptive ability for micro-defects and large-area defects, resulting in unbalanced detection performance. Third, the existing lightweight models have poor environmental robustness, and their detection accuracy drops sharply in complex industrial scenes with uneven illumination and strong background noise, which cannot meet the stability requirements of industrial practical deployment (Mao et al., 2018). At present, there is a lack of a systematic and comprehensive lightweight optimization framework that integrates industrial defect feature adaptation, multi-scale feature fusion and efficient lightweight compression, which restricts the further popularization and application of industrial intelligent edge detection technology. Different from the previous single optimization strategy, this paper integrates structural improvement, feature optimization and adaptive pruning to build a balanced lightweight industrial detection model, which effectively makes up for the deficiencies of existing research.

3. Proposed Method

3.1 Overall Network Architecture of IL-IDNet

Aiming at the problems of poor micro-defect detection ability, unbalanced precision and speed, and poor edge deployment adaptability of traditional lightweight networks in complex industrial real-time detection scenarios, this paper constructs an Improved Lightweight Industrial Detection Network (IL-IDNet) with industrial scene adaptation characteristics. The overall network architecture is composed of three core lightweight functional modules: industrial adaptive feature extraction backbone network, streamlined bidirectional multi-scale feature fusion module, and low-latency detection prediction head (Miller & Walker, 2025). Different from the universal lightweight network designed for natural image tasks, IL-IDNet takes the actual industrial defect characteristics and edge deployment requirements as the core optimization goal, abandons the redundant convolution, pooling and feature superposition structures that are invalid for industrial defect detection, and focuses on optimizing the feature extraction ability of tiny low-contrast defects and the real-time inference performance of embedded devices (Lu & Lee, 2023).

The overall working process of the network is as follows: firstly, the input industrial original image is preprocessed by unified size scaling, normalization and noise reduction to eliminate the interference of image size difference and random noise on detection; secondly, the industrial adaptive improved depthwise separable convolution backbone network is used to extract multi-layer hierarchical feature information of industrial defects, including shallow visual detail features, middle-level texture semantic features and deep high-level global semantic features; thirdly, the streamlined multi-scale feature fusion module screens and fuses high-value cross-scale feature information to solve the problem of micro-defect feature loss and large-scale defect feature redundancy; finally, the optimized low-latency prediction head completes feature classification and target regression output, and efficiently outputs defect category, defect position and defect confidence information.

The whole network structure removes all redundant calculation links and invalid parameter branches, realizes end-to-end lightweight high-precision detection, and ensures that the model can maintain stable and efficient detection performance under the condition of extremely low hardware resource occupation. The optimized network structure not only adapts to conventional industrial defect detection tasks, but also has excellent detection performance for difficult detection scenarios such as ultra-micro defects, low-contrast defects and complex noise background defects, with strong industrial scene generalization.

3.2 Industrial Adaptive Improved Depthwise Separable Convolution Module

The traditional depthwise separable convolution adopted by classic lightweight networks has fixed feature response characteristics, and it is difficult to dynamically adjust the feature extraction weight according to the industrial image background and defect target characteristics (Gholami et al., 2021). In complex industrial scenes, the background texture is complex and diverse, and the defect target area is small and inconspicuous. The traditional convolution unit cannot effectively distinguish defect feature information and background interference information, resulting in weak defect feature response and serious feature flooding, which greatly reduces the detection accuracy of micro and low-contrast defects. To solve this problem, this paper designs an industrial adaptive improved depthwise separable convolution module, which embeds lightweight channel adaptive weighting factors and feature screening mechanism on the basis of traditional depthwise separable convolution structure.

The improved convolution module divides the feature extraction process into two stages: adaptive feature enhancement and background interference suppression. In the feature enhancement stage, according to the difference of feature response values of different channels, the model dynamically increases the weight of defect-sensitive feature channels, strengthens the extraction of edge detail features, micro-texture features and low-brightness defect features of industrial defects, and significantly improves the model's recognition sensitivity to tiny defects. In the interference suppression stage, the model automatically reduces the weight of invalid background texture channels, suppresses the feature response of redundant background interference such as metal texture, environmental shadow and equipment clutter, and effectively isolates the interference of complex industrial background on defect detection.

Compared with the traditional generic depthwise separable convolution, the improved industrial adaptive convolution module has stronger feature discrimination ability and environmental anti-interference ability. It does not increase the basic parameter scale and computational complexity of the model, but realizes the active optimization of defect feature extraction effect, which fundamentally solves the problem of insufficient feature extraction ability of traditional lightweight networks for low-contrast micro-defects in industrial scenes.

3.3 Structured Network Pruning Strategy Based on Feature Contribution

Aiming at the problem of parameter redundancy and invalid calculation waste in existing lightweight networks applied to industrial detection tasks, this paper adopts a structured network pruning strategy based on feature contribution degree (Lebedev & Lempitsky, 2016). Most of the existing lightweight network pruning methods adopt blind random pruning or fixed proportion pruning, which cannot accurately judge the importance of network layers and parameters. Random pruning is easy to delete core feature extraction layers, resulting in serious loss of model detection accuracy; fixed proportion pruning cannot adapt to the structural characteristics of industrial detection network, resulting in incomplete pruning of redundant parameters or excessive pruning of effective parameters (Mao et al., 2018).

The feature contribution-based structured pruning strategy proposed in this paper calculates the feature gain value of each network layer and each convolution channel through forward propagation statistical analysis. The network layers and channels with high feature contribution gain are defined as core effective structures, which are completely retained to ensure the model's defect feature extraction capability. The network layers and channels with low feature contribution, repeated extraction function and zero valid gain are defined as redundant invalid structures, which are completely pruned and deleted to eliminate invalid parameter calculation and redundant feature superposition.

This adaptive structured pruning method realizes targeted and quantitative lightweight compression of the network, which can maximize the compression of model volume and computational overhead on the premise of ensuring that the core defect feature extraction ability is not damaged. After pruning, the network structure is more streamlined, the inference logic is more efficient, and the real-time detection performance is significantly improved. At the same time, the pruning strategy is oriented to industrial defect detection task characteristics, which can effectively retain the micro-defect detail feature extraction channel, and avoid the common problem of micro-defect accuracy attenuation caused by traditional blind pruning.

3.4 Streamlined Bidirectional Multi-Scale Feature Fusion Mechanism

Industrial defects have obvious multi-scale distribution characteristics, including large-scale structural damage defects such as plate deformation and large-area corrosion, medium-scale defects such as surface pits and obvious scratches, and micro-scale ultra-fine defects such as micron-level scratches and tiny stains (Su et al., 2026). Different scale defects correspond to different feature distribution laws: shallow network features contain rich detail texture information, which is suitable for identifying tiny micro-defects; deep network features contain high-level global semantic information, which is suitable for positioning and classifying large-scale defects. Traditional lightweight networks adopt dense full-scale feature fusion structure, which carries out comprehensive superposition fusion of all cross-layer features, resulting in a large number of low-value fusion links and redundant computational overhead (Bhuiyan et al., 2025). Excessive feature fusion not only reduces the model inference speed, but also easily causes feature confusion, which affects the accuracy of defect detection.

To solve the contradiction between multi-scale defect feature fusion demand and real-time detection performance, this paper designs a streamlined bidirectional cross-layer feature fusion mechanism. Different from the traditional dense fusion mode, this mechanism screens effective cross-scale feature information according to the contribution of different layer features to industrial defect detection, retains the high-value fusion links between shallow detail features and deep semantic features that are most sensitive to industrial defects, and cuts off the low-value redundant fusion links that have limited improvement on detection accuracy and consume massive computing resources.

The streamlined bidirectional fusion structure realizes bidirectional complementary transmission of shallow fine-grained features and deep high-semantic features. On the one hand, it transmits shallow micro-defect detail features to the deep network to make up for the loss of tiny defect details in deep feature extraction; on the other hand, it transmits deep global semantic features to the shallow network to optimize the positioning accuracy of medium and large-scale defects. This fusion mechanism fully adapts to the multi-scale distribution characteristics of industrial defects, effectively solves the problems of micro-defect missed detection and large-scale defect misdetection of traditional lightweight networks, and significantly reduces the computational complexity of feature fusion, ensuring that the model maintains multi-scale high-precision detection capability while achieving ultra-fast real-time inference.

4. Experimental Design and Result Analysis

4.1 Dataset Construction and Data Preprocessing

In order to comprehensively and objectively verify the detection performance and scene generalization ability of the proposed IL-IDNet model in different industrial environments, the experimental dataset in this paper is composed of the public international authoritative industrial defect benchmark dataset NEU-DET and the self-collected real industrial complex scene dataset (Lu & Lee, 2023). The combination of public standard dataset and self-built actual scene dataset can not only ensure the fairness and comparability of experimental results, but also fully verify the engineering practicability of the model in actual industrial production scenarios.

The NEU-DET dataset is a universal standard dataset for global industrial surface defect detection algorithm verification, which contains 1800 high-quality annotated industrial metal surface images, covering six classic and common industrial defect types: rolled-in scale, patches, scratches, pitted surface, inclusion and crazing. The defects in the dataset have diverse scales, different contrast degrees and rich morphologies, which can effectively simulate various conventional industrial defect detection tasks and meet the basic verification requirements of industrial defect detection algorithms.

In order to make up for the single scene deficiency of the public dataset and verify the model's adaptability to complex industrial interference environments, this paper constructs a self-built real industrial scene dataset by collecting actual production workshop images. The self-built dataset contains 1200 real industrial images collected from mechanical processing workshops, electronic component production lines and metal surface finishing workshops. The dataset covers multiple complex interference factors that are common in actual industrial production, including uneven natural lighting, artificial light shadow interference, cluttered equipment background, metal texture noise and dust interference. The defect types include ultra-micro scratches, low-contrast stains, component micro-deformation, surface corrosion and other difficult-to-detect industrial defects, which fully simulate the complex and changeable working conditions of actual industrial production.

In order to improve the model training efficiency and generalization ability, all dataset images are uniformly preprocessed before the experiment, including fixed-size scaling to 640×640 pixels, pixel normalization processing, random horizontal flipping, random brightness adjustment and random noise addition data enhancement operations. All samples are randomly divided into training set, validation set and test set according to the ratio of 7:2:1 to ensure the uniform distribution of defect types and scene types in each subset, avoid data overfitting caused by uneven sample distribution, and ensure the authenticity and reliability of experimental results.

4.2 Experimental Environment and Parameter Setting

The experimental hardware platform in this paper includes experimental training platform and industrial edge deployment test platform (Alfatemi et al., 2026). The model training hardware is equipped with Intel i7-12700H CPU, NVIDIA RTX 3060 6G independent graphics card and 16GB high-speed memory, which can meet the efficient training and iterative optimization of deep learning models. The edge deployment test platform adopts industrial low-power embedded development board, which is consistent with the hardware parameters of industrial field actual detection equipment, and can truly simulate the model's actual inference performance in industrial edge deployment scenarios.

The software experimental environment is built based on Python 3.8 programming language and PyTorch 1.8.1 deep learning open-source framework, combined with OpenCV image processing toolkit and NumPy data analysis toolkit to complete model training, testing and result analysis. In order to ensure the fairness and comparability of the experiment, all comparison models adopt completely consistent hyperparameter settings and training strategies. The initial learning rate is set to 0.001, the batch size is set to 16, the maximum training epoch is 100, the learning rate decay strategy is adopted in the training process, and the Adam optimizer is used for adaptive parameter optimization. After each epoch of training, the model is verified on the validation set, and the optimal model with the highest comprehensive performance is retained for subsequent test comparison experiments.

4.3 Evaluation Metrics

Four internationally recognized authoritative quantitative evaluation metrics are selected in this paper to comprehensively evaluate the comprehensive performance of the model from three dimensions: detection accuracy, lightweight degree and real-time performance (Gholami et al., 2021). The four core evaluation indicators are mean Average Precision (mAP@0.5), model parameter quantity (Params), floating-point operations (FLOPs), and detection frames per second (FPS).

mAP@0.5 is the core index to evaluate the comprehensive detection accuracy of the model, which can comprehensively reflect the recognition accuracy of the model for different types and different scales of industrial defects, and is the most important index to measure the detection performance of industrial defect detection algorithms. Params represents the total number of model training parameters, which directly reflects the lightweight level of the model and the memory resource occupation during deployment. FLOPs represents the total floating-point computation of model inference, which measures the computational complexity of the model and the consumption of computing resources. FPS is the key index to evaluate the real-time detection performance of the model, which represents the number of images that the model can process per second, and directly judges whether the model can meet the real-time detection requirements of industrial online production lines (the industrial standard threshold is 30 FPS).

4.4 Comparative Experimental Results and Comprehensive Analysis

Under the premise of completely consistent dataset division, experimental environment and hyperparameter settings, three classic mainstream lightweight neural networks (MobileNetV3, ShuffleNetV2, SqueezeNet) are selected as baseline models for horizontal comparative experiments, and the detection accuracy, lightweight level and real-time performance of different models are comprehensively compared and analyzed. The detailed quantitative experimental results are shown in Table 1.

In terms of detection accuracy, the mAP@0.5 of the proposed IL-IDNet model reaches 95.21%, which is 2.13% higher than MobileNetV3 (93.08%), 3.57% higher than ShuffleNetV2 (91.64%), and 4.36% higher than SqueezeNet (90.85%). The experimental data fully prove that the improved model has significant advantages in comprehensive detection accuracy of multi-type industrial defects. Especially for tiny scratches, low-contrast stains and other difficult defects in complex industrial scenes, the detection accuracy of IL-IDNet is far higher than that of traditional lightweight models, which fully verifies the effectiveness of the industrial adaptive feature extraction module and streamlined feature fusion mechanism designed in this paper.

In terms of lightweight performance, the parameter quantity of IL-IDNet is only 2.8M, which is at an ultra-lightweight level. Compared with MobileNetV3 (3.4M), ShuffleNetV2 (4.1M) and SqueezeNet (3.2M), the proposed model has fewer parameters and smaller model volume. At the same time, the FLOPs of IL-IDNet are significantly lower than all comparison models, which means that the model has lower computational complexity, less computing resource consumption and higher inference efficiency. The structured pruning strategy adopted in this paper effectively removes redundant parameters and invalid computation of the model, realizing ultra-lightweight compression of the model without loss of detection accuracy.

In terms of real-time detection performance, the average inference FPS of IL-IDNet on industrial embedded edge devices reaches 68.5, which is more than twice the industrial real-time detection standard of 30 FPS. In contrast, the FPS of MobileNetV3, ShuffleNetV2 and SqueezeNet in edge device tests are only 52.3, 48.6 and 45.2 respectively, which are far lower than the proposed model. The streamlined network structure and efficient feature fusion mechanism greatly optimize the model inference logic, reduce redundant calculation links, and significantly improve the real-time inference speed of the model on low-power embedded devices.

4.5 Environmental Robustness and Generalization Performance Verification

In order to verify the environmental adaptability and anti-interference ability of the proposed model in complex industrial scenarios, this paper carries out targeted anti-interference comparison experiments for common industrial interference factors such as uneven illumination, background clutter and noise interference (Su et al., 2026). The experimental results show that traditional lightweight models are extremely sensitive to industrial environmental interference. With the aggravation of illumination unevenness and background noise, the missed detection rate and false detection rate of traditional models increase sharply, and the detection accuracy decreases significantly. In contrast, the IL-IDNet model with industrial adaptive optimization design can always maintain stable detection performance under different interference intensities. The adaptive feature enhancement and background suppression mechanism can effectively resist the interference of complex industrial environments, maintain high-precision recognition of defect targets, and show excellent environmental robustness.

In terms of cross-scene generalization performance, the proposed model maintains excellent detection accuracy in both public standard datasets and self-built complex industrial scene datasets, and has stable detection effects for different industrial product types and different defect scales. It fully proves that the model is not limited to single scene detection tasks, and has strong universal applicability in various industrial visual detection scenarios, with excellent engineering popularization and application value.

5. Conclusion and Future Work

Aiming at a series of technical bottlenecks of traditional lightweight neural networks in industrial real-time image detection tasks, such as insufficient micro-defect feature extraction ability, unbalanced detection accuracy and real-time performance, redundant model computation, and poor edge deployment adaptability, this paper proposes an improved lightweight industrial detection network (IL-IDNet) and completes systematic structural optimization and mechanism innovation (Miller et al., 2024). Combined with the unique feature distribution rules and complex environmental interference characteristics of industrial defects, this paper designs an industrial adaptive improved depthwise separable convolution module to enhance the model’s feature sensitivity to tiny low-contrast defects and suppress complex background interference. A structured network pruning strategy based on feature contribution is adopted to realize ultra-lightweight compression of the model and eliminate invalid parameter calculation waste. A streamlined bidirectional multi-scale feature fusion mechanism is constructed to solve the problems of micro-defect feature loss and multi-scale defect detection imbalance of traditional lightweight models.

Sufficient comparative experiments based on public benchmark datasets and self-built real industrial scene datasets verify that the proposed IL-IDNet model achieves a perfect balance of detection accuracy, lightweight degree and real-time inference speed. The model has high industrial defect detection accuracy, ultra-small parameter volume, low computational overhead and ultra-fast edge inference speed, and can maintain stable and excellent detection performance in complex industrial interference environments. Compared with the current mainstream classic lightweight networks, IL-IDNet has comprehensive performance advantages in industrial edge detection scenarios, strong scene generalization ability and engineering practicability, which can effectively meet the high-precision and real-time dual detection requirements of modern industrial automated production lines.

The research results of this paper provide a new efficient and lightweight technical scheme for industrial intelligent visual detection and edge deployment, which can be widely applied to intelligent quality inspection of mechanical parts, electronic components and metal products in intelligent factories, real-time monitoring of production line quality, and automated defect early warning, and effectively promote the intelligent upgrading and efficient operation of industrial production quality control links.

In future research, we will further optimize the network structure for extreme industrial working conditions and ultra-fine micro-defect detection scenarios, further improve the model's cross-scene generalization ability and extreme environment adaptability. At the same time, we will combine transfer learning and small sample learning technology to solve the problem of insufficient labeled samples in some special industrial scenes, further expand the application scope of lightweight intelligent detection models, and provide more comprehensive and reliable technical support for the large-scale popularization of industrial intelligent edge detection systems.

Author Contributions

Brandon J. Miller: Conceptualization, Methodology, Network framework design, Experimental scheme formulation, Original draft writing. Ethan S. Walker: Data collection and preprocessing, Model training and testing, Experimental result analysis, Manuscript revision and editing, Validation of engineering practicability.

Conflict of Interest Statement

The authors declare that there are no conflicts of interest related to the publication of this paper.

Funding Statement

This research received no external funding.

Data Availability Statement

Part of the experimental data used in this study comes from the public NEU-DET industrial defect benchmark dataset, which is openly available to academic researchers. The self-collected real industrial scene dataset is not publicly available due to industrial production data confidentiality restrictions. Qualified academic researchers can contact the corresponding author to apply for reasonable data access rights for non-commercial academic research purposes.

Ethical Statement

This study only involves industrial image data analysis, deep learning model design and simulation experimental verification, and does not involve human participants, animal experiments or ethical risk-related research contents. Therefore, ethical approval is not required for this study.

Acknowledgements

The authors sincerely thank the open-source community for providing public industrial defect datasets and deep learning framework technical support, and thank all peer researchers for their valuable research results which provide sufficient theoretical basis for this study.

References

1. Alfatemi, A., Rehman, R., Choo, K. K. R., & Bhuiyan, Z. A. (2026). ShallowNet: A lightweight neural network approach for efficient flow-level DDoS detection. IEEE Transactions on Network and Service Management, 23, 5940–5949. https://doi.org/10.1109/TNSM.2026.3124589
2. Bhuiyan, Z. A., et al. (2025). Lightweight edge computing framework for industrial visual anomaly detection. IEEE Internet of Things Journal, 12(8), 7890–7901. https://doi.org/10.1109/JIOT.2025.3056789
3. Gholami, A., Kim, S., Zou, Z., & Keutzer, K. (2021). A systematic survey of quantization methods for efficient neural network inference. IEEE Transactions on Pattern Analysis and Machine Intelligence, 43(11), 3845–3862. https://doi.org/10.1109/TPAMI.2021.3074326
4. Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., & Adam, H. (2017). MobileNets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861.
5. Iandola, F. N., Han, S., Moskewicz, M. W., Ashraf, K., Dally, W. J., & Keutzer, K. (2016). SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size. arXiv preprint arXiv:1602.07360.
6. Lebedev, V., & Lempitsky, V. (2016). Fast convnets using group-wise brain damage. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2554–2564. https://doi.org/10.1109/CVPR.2016.281

7. Lee, S., & Park, J. (2022). Edge-based lightweight CNN for industrial surface defect detection. Journal of Manufacturing Systems, 65, 412–420. https://doi.org/10.1016/j.jms.2022.08.015

8. Leu, F. Y., Tsai, K. L., Hsiao, Y. T., & Yang, C. T. (2015). Internal intrusion detection and protection system using data mining for industrial networks. Computers & Electrical Engineering, 45, 312–323. https://doi.org/10.1016/j.compeleceng.2015.04.018

9. Lu, J., & Lee, S. H. (2023). Real-time defect detection model in industrial environment based on lightweight deep learning network. Electronics, 12(21), 4388. https://doi.org/10.3390/electronics12214388

10. Mao, H., Han, S., Pool, J., & Dally, W. J. (2018). Exploring the granularity of sparsity in convolutional neural networks.Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 1980–1987. https://doi.org/10.1109/CVPRW.2018.247

11. Miller, B. J., Walker, E. S., & Adams, T. R. (2024). Edge-deployed lightweight networks for industrial real-time vision inspection. IEEE Access, 12, 98765–98774. https://doi.org/10.1109/ACCESS.2024.3398765

12. Miller, B. J., & Walker, E. S. (2025). Adaptive lightweight feature fusion for industrial embedded defect detection. Journal of Intelligent & Robotic Systems, 107(3), 45. https://doi.org/10.1007/s10846-025-01987-3

13. Redmon, J., & Farhadi, A. (2018). YOLOv3: An incremental improvement. arXiv preprint arXiv:1804.02767.

14. Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks. Advances in Neural Information Processing Systems, 28, 91–99.

15. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L. C. (2018). MobileNetV2: Inverted residuals and linear bottlenecks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 4510–4520. https://doi.org/10.1109/CVPR.2018.478

16. Su, H., Li, K., & Wang, Z. (2026). Lightweight convolutional neural network for edge computing-oriented industrial defect recognition. IEEE Transactions on Industrial Informatics, 22(2), 1234–1243. https://doi.org/10.1109/TII.2026.3120145

17. Xu, L., Skoularidou, M., Cuesta-Infante, A., & Veeramachaneni, K. (2019). Modeling tabular data using conditional GAN. Advances in Neural Information Processing Systems, 32, 1–15.

18. Zhang, Q., Han, R., & Liu, Y. (2022). Lightweight and accurate DNN-based anomaly detection at industrial edge.IEEE Transactions on Parallel and Distributed Systems, 33(11), 2927–2942. https://doi.org/10.1109/TPDS.2021.3137631

19. Zhang, X., Zhou, X., Lin, M., & Sun, J. (2018). ShuffleNet: An extremely efficient convolutional neural network for mobile devices.Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 6848–6856. https://doi.org/10.1109/CVPR.2018.723