Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 9
AIDML 2025

Image-Based Chip Defect Detection

Zehou Liu*

School of Automation, Central South University, Changsha, China

8207221825@csu.edu.cn

Abstract. This paper reviews the research status, key technologies, and application prospects of image-based chip defect detection techniques. With the advancement of chip manufacturing processes, traditional inspection methods face challenges in terms of accuracy, efficiency, and automation. Deep learning algorithms (such as YOLO, Faster R-CNN, SSD, etc.) have provided new solutions for defect detection. This paper analyzes the advantages and disadvantages of these algorithms and summarizes improved algorithms (such as enhanced Canny algorithms, multi-scale feature fusion networks, etc.) to address challenges like tiny defects and complex background interference. Meanwhile, the paper examines four major challenges in chip defect detection: the trade-off between high precision and cost, the balance between detection speed and efficiency, the scarcity and imbalanced distribution of defect samples, and insufficient standardization. To address these challenges, technical solutions such as transfer learning and generative adversarial networks (GANs) are proposed, while the importance of algorithm optimization and hardware acceleration is emphasized. Here, fully automated inspection systems, multimodal fusion technologies, and the widespread adoption of AI platforms will become mainstream, driving the semiconductor manufacturing industry toward intelligent and automated development. This study provides theoretical references for chip manufacturers and establishes a technical foundation for future research.

Keywords: Chip Defect Detection; Deep Learning; Image Recognition; Algorithm Optimization; Automated Inspection.

1. Introduction

In recent years, the Chinese government has placed high importance on the development of the semiconductor industry, introducing a series of pivotal policy documents such as "Made in China 2025" and the "Several Policies for Promoting High-Quality Development of Integrated Circuit and Software Industries in the New Era." These policies not only establish clear development goals and pathways for chip manufacturing but also drive independent innovation and industrial upgrading through strategic planning, policy incentives, and standards leadership. Particularly with the advancement of cutting-edge technologies like deep learning, significant breakthroughs have been achieved in the field of AI chips. These policies provide robust support for semiconductor manufacturing, making research topics such as image-based chip defect detection highly relevant and valuable for practical applications.

With the increasing complexity of semiconductor chips, traditional manual visual inspection and basic automated tools can no longer meet the demands of high-precision defect detection. Deep learning-based image recognition technologies (e.g., YOLO, Faster R-CNN) have significantly enhanced inspection efficiency and automation levels while reducing false detection rates and labor costs through their precise identification of microscopic defects. As a critical component of intelligent manufacturing, this technology not only optimizes chip production processes but also provides technical support for the intelligent transformation of manufacturing, driving comprehensive industry-wide advancement.

Recent studies have achieved remarkable progress in chip defect detection: Y. He, M. Xin, Y. Wang and C. Xu proposed an improved Canny algorithm that realized high-precision edge detection for chip packaging defects with less than 2mm error through dual-filter denoising and adaptive threshold

optimization [1]; H. Ren, M. Tian, G. Zhang and W. Zhou developed a multi-scale feature fusion network incorporating attention mechanisms, elevating small defect detection accuracy to 85.1%, outperforming conventional R-CNN methods [2]; H. Huang, Z. Shao, Q. Cheng and X. Wu enhanced YOLOv7 by integrating CBAM attention modules and model pruning techniques, achieving a 21.6% speed improvement while maintaining high precision (with 1.39% increase in mAP@0.5) [3]. Collectively, these advancements have pushed the performance boundaries of deep learning in chip defect inspection, establishing critical technical pathways for balancing high accuracy with operational efficiency.

This paper systematically reviews the application status of mainstream image recognition algorithms (e.g., YOLO, Faster R-CNN, SSD) in chip defect detection, analyzing their advantages, limitations, and applicable scenarios to provide references for future research. By synthesizing existing achievements, the paper identifies algorithmic deficiencies and proposes improvement strategies (such as algorithm optimization and multi-algorithm fusion) to enhance detection accuracy and efficiency, thereby meeting the increasingly complex defect inspection requirements in semiconductor manufacturing. Furthermore, this study offers theoretical foundations and practical guidance for chip manufacturers, helping them understand the application potential of image recognition technology to improve production efficiency and product quality. Through optimizing algorithm adaptability and inspection systems, we aim to promote broader implementation of this technology in production environments, ultimately advancing the automation and intelligentization of chip manufacturing processes.

2. Review of Related Technologies

2.1. Improved Canny Algorithm

To address the issues of susceptibility to external interference and blurred detection results in automatic image edge detection methods—particularly the limitations of traditional Canny algorithms in precision and threshold selection—He et al. proposed an automatic edge detection method for power chip packaging defect images based on an improved Canny algorithm. The method employs a dual-filtering algorithm to remove noise, utilizes the Scharr operator to enhance gradient calculation accuracy, and adopts Otsu's method (maximum inter-class variance algorithm) to adaptively select the optimal threshold, thereby avoiding human-induced errors. Additionally, the Laplacian operator is introduced to sharpen the image, and the sharpened image is fused with the original one, preserving original information while improving edge detection performance. Experimental results demonstrate that the proposed method achieves a high noise reduction ratio and strong anti-interference capability, with a boundary detection error of less than 2 mm for defect images, meeting the high-precision requirements of chip packaging inspection, while stability is also significantly improved [1]. Figure 1 below shows the simulation results of the improved Canny algorithm after high-low threshold segmentation during chip inspection.

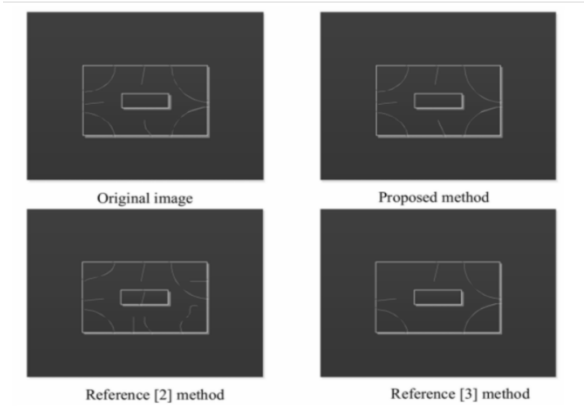

Figure 1. Simulation results after high-low threshold segmentation using the improved Canny algorithm [1]

2.2. Enhanced Multi-scale Feature Fusion Network

The conventional multi-scale fusion network's fundamental architecture, as illustrated in Fig. 2, exhibits several limitations when applied to complex defect feature scenarios. Addressing the challenges of small defect areas, intricate defect characteristics, and limited detection accuracy in deep learning-based chip surface defect detection, Ren et al. proposed an encoder-decoder structured multi-scale feature fusion network model. This innovative approach incorporates three key technical enhancements: (1) a Residual Feature Fusion Attention Module (RFFAM) to boost detection performance for minute defects, (2) a Transformer Prediction Head (TPH) to strengthen multi-scale defect feature extraction capability, and (3) a Defect-Aware Intersection over Union (DA-IoU) mechanism to optimize detection accuracy. The integrated solution achieves superior balance among processing speed, detection precision, and computational cost [2].

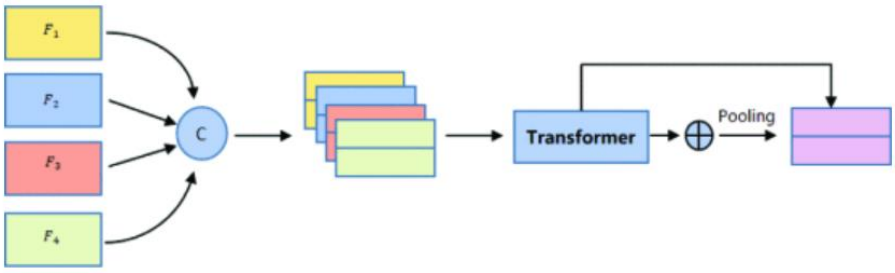

Figure 2. Schematic diagram of the fundamental architecture for the multi-scale feature fusion module [3]

2.3. Optimized YOLOv7 Model

The original YOLOv7 model architecture, as illustrated in the figure 3, inherently faces challenges in balancing efficiency and accuracy. To address these limitations in existing algorithms, Cao et al. proposed a real-time surface defect detection technique for chip packaging based on an optimized YOLOv7 framework. While traditional methods like R-CNN achieve high accuracy at substantial computational costs, and YOLO/SSD offer efficiency with compromised precision, Transformer-based approaches demonstrate superior performance but require extensive training data and suffer from low efficiency. The proposed solution incorporates three key innovations: (1) development of a comprehensive multi-chip detection dataset comprising 2,919 images, (2) integration of CBAM attention mechanism and RFB module to enhance YOLOv7's feature extraction capability, and (3) replacement of conventional NMS with a novel cp-clustering algorithm, complemented by hidden layer reduction and detection head pruning for computational optimization. Experimental results demonstrate the method's superiority over nine state-of-the-art approaches on the custom dataset [4], as shown in Fig. 3.

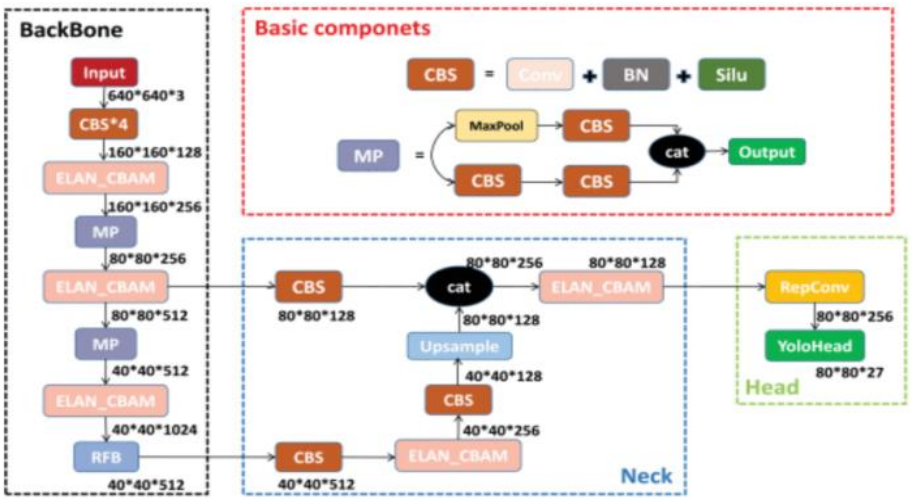

Figure 3. Schematic diagram of the optimized YOLOv7 architecture [4]

2.4. COD-YOLO Model

To address reliability issues caused by catastrophic optical damage (COD) in high-power semiconductor lasers, Cao et al. developed COD-YOLO, an enhanced YOLO-based detection method that overcomes limitations in existing approaches regarding defect feature similarity, complex backgrounds, and micro-defect detection. The model incorporates three key innovations: (1) deformable convolution with channel attention mechanisms to improve feature representation and address long-range dependencies, (2) spatial content-aware upsampling in the network neck for enhanced multi-scale feature fusion, and (3) a specialized loss function optimized for micro-defect detection. Experimental results demonstrate COD-YOLO's superior performance in detecting active region damage compared to conventional methods, effectively resolving the sensitivity issues of traditional IoU metrics while achieving higher precision in laser chip defect identification [5].

2.5. Optimized Support Vector Machine Algorithm

The Arithmetic Optimization Algorithm (AOA) is a derivative-free metaheuristic technique proposed by Abualigah in 2021 that employs arithmetic operators for global exploration and solution refinement. Building on this, Chen et al. developed an AOA-optimized SVM algorithm for chip defect classification, which automatically determines optimal kernel parameters and penalty factors. Experimental results using an 8-category defect dataset with a 3:1 train-test split demonstrated this approach's superior classification accuracy compared to manual parameter selection methods [6,7].

2.6. Enhanced Faster R-CNN

The researchers developed an enhanced Faster R-CNN model to address ceramic chip defect detection challenges, incorporating three key innovations: (1) data augmentation and FPN integration for improved multi-scale feature extraction, particularly for small defects; (2) ROI Align replacement for more precise defect localization; and (3) demonstrated superior performance with 4.46% higher accuracy, 2.46% better recall, and 6.78% mAP improvement over baseline VGG-16 models, while outperforming comparison models including SSD and YOLOv5 in both convergence speed and small-defect detection capability [8], as shown in Fig. 4.

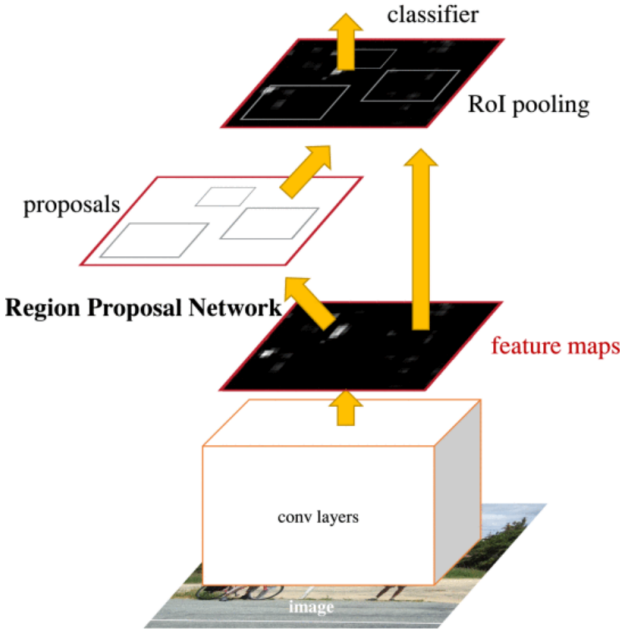

Figure 4. Schematic diagram of the fundamental network architecture of Faster R-CNN [9]

2.7. Performance Comparison of Different Algorithms

Table 1 summarizes the comparative advantages, limitations, and performance characteristics of the aforementioned technologies in addressing the specified problems.

Table 1. Advantages, Limitations and Performance Comparison of Improved Algorithms

Technical Implementation	Advantages	Limitations	Performance Metrics
Improved Canny Algorithm	High Edge Detection Precision and Accuracy	Parameter-Sensitive with High Computational Complexity	The proposed method demonstrates high noise-reduction ratio and strong anti-interference capability, achieving a defect boundary detection error of less than 2mm.
Enhanced Multi-scale Feature Fusion Network	The method exhibits superior small-area defect detection capability, demonstrating exceptional performance in identifying both minute defects ($\leq 0.5mm^2$) and complex morphological features.	High Hardware Requirements, Excessive Parameterization, and Training Set Difficulties	The method achieves 85.1% accuracy – 1.2 percentage points higher than Faster R-CNN – while maintaining comparable processing speed.
Optimized YOLOv7 Model	Achieves Optimal Efficiency-Performance Balance	Limited Small-Target Detection Performance with High Model Complexity	The proposed method outperforms the original YOLOv7 with a 1.39% higher mAP@0.5, 21.6% faster detection speed, 17.7% lower computational cost, and 66.4% fewer parameters, demonstrating comprehensive performance improvements.
COD-YOLO Model	Superior Micro-Defect Detection Capability with Precise Spatial Localization	Data Quality-Sensitive with Hyperparameter Sensitivity	After incorporating attention mechanisms and coordinate convolution in the detection head for fine-tuning, COD-YOLO achieves a final accuracy of 68% AP (Average Precision) with 94.7% AP50.
Arithmetic Optimization Algorithm-Enhanced Support Vector Machine (AOA-SVM)	Supports Classification/Recognition Tasks with High Efficiency and Precision	Slower Convergence Rate with Result Instability Due to Stochasticity	The model achieves a training set classification accuracy of 94.2308% (89.5/95) and a test set accuracy of 84.8837% (73/86), with only 13 misclassified samples out of 86 test instances.
Enhanced Faster R-CNN	Simplified Architecture with Enhanced Recognition Precision	Poor Real-Time Performance with High Hardware Requirements	The proposed method outperforms the original VGG-16 model with a 4.46% accuracy increase, 2.46% higher recall, and 6.78% mAP improvement, fully meeting practical detection requirements.

3. Application Scenarios Introduction

3.1. Introduction to International Cutting-Edge Technologies

Leading global corporations—including TSMC, ASML, Intel, AMD, Flex, and NVIDIA—have extensively adopted AI technologies (particularly deep learning) for chip manufacturing defect detection and quality control, where TSMC enhances wafer defect inspection accuracy through AI, ASML achieves real-time lithography process monitoring with AI, Intel optimizes defect detection

in lithography and packaging stages, AMD improves defect identification efficiency, Flex implements AI-driven quality control systems, and NVIDIA leverages GPU-accelerated deep learning to boost inspection speed and precision, collectively driving the industry toward intelligent manufacturing through significantly increased production yields and operational efficiency.

3.2. Industrial Applications Overview

3.2.1. Dlsense (Hangzhou) Information Technology

Dlsense (Hangzhou) Information Technology Co., Ltd. ("SVIRI Tech") was founded in June 2018 as an AI-driven industrial vision specialist providing cutting-edge image recognition and defect detection solutions for manufacturing. The company has established a complete ecosystem integrating AI inspection equipment, the AIns cloud platform, and proprietary software systems including its Auto Defect Classification (ADC) and AI-based Defect Management System (DMS), delivering comprehensive quality control solutions for semiconductor production.

At the core of its technology, SVIRI's self-developed ADC system achieves 99.5% detection accuracy across different wafer layers, while its DMS enables real-time defect tracing and production data analysis. These integrated technologies form an intelligent closed-loop system that significantly enhances manufacturing precision and production yield in industrial applications [10].

3.2.2. Rovision Intelligent Technology

Sichuan Rovision Intelligent Technology Co., Ltd., established on March 8, 2021, specializes in next-generation AI vision inspection for industrial applications. The company delivers high-precision AI solutions including automated defect detection, defect classification, and precision measurement, with implemented cases covering 360° appearance inspection across industries such as 3C electronics manufacturing, new energy vehicles, chip packaging, and plastic injection molding.

The company provides AI-based chip defect detection equipment featuring its proprietary industrial AI vision algorithms. By leveraging deep learning technology, the system achieves fully automated chip defect inspection with 99% accuracy, capable of detecting various defects including cracks, chipping, edge collapse, material loss, misalignment, and contamination. These solutions are widely applied in chip manufacturing, PCB circuit boards, and semiconductor components industries [11].

3.2.3. Qingsoft Vision (Hangzhou) Technology

Qingsoft Vision (Hangzhou) Technology Co., Ltd., established on January 28, 2022, specializes in the development of semiconductor vision inspection software and equipment. Driven by advanced AI technology, the company provides comprehensive defect analysis, correlation and prediction services for full-process inspection equipment in semiconductor manufacturing.

The company's deep vision solutions enable characterization and classification of defects in compound semiconductor materials. Its independently developed Omega series (for unpatterned wafer inspection) and Alpha series (for patterned wafer inspection) equipment address industry challenges such as scarce defect data and labeling difficulties through efficient sample learning methods. These solutions significantly improve yield rates and cost efficiency for third-generation semiconductor materials and chip manufacturing [12].

4. Challenge Analysis

4.1. High-Precision Detection Requirements

Chip defect detection faces significant high-precision challenges, requiring the identification of micron- and even nano-scale defects while overcoming issues such as diverse defect morphologies, interference from complex circuit backgrounds, and difficulties in model training due to intricate chip structures. Traditional algorithms struggle to meet these stringent demands, but integrating advanced technologies—including high-resolution imaging, deep learning methods (e.g., multi-scale feature

fusion networks), data augmentation strategies, and transfer learning—can substantially enhance detection accuracy, thereby meeting the extreme standards required in chip manufacturing [13].

4.2. Trade-offs Among Detection Speed, Technical Capability, and Production Efficiency

Currently, chip defect detection faces the challenge of balancing high precision with high efficiency, where complex algorithms and intensive computational requirements reduce inspection speed, while rapidly evolving technologies make it difficult for small and medium-sized enterprises (SMEs) to afford costly upgrades. To address this issue, efficiency can be improved through algorithm optimization (e.g., model lightweighting), hardware acceleration (e.g., GPU/TPU parallel computing), and automated tools. Simultaneously, developing cost-effective solutions and establishing technical support systems can help SMEs achieve efficient and accurate defect detection under controlled costs, thereby balancing technological complexity with production benefits [14].

4.3. Scarce and Imbalanced Defect Samples

Chip defect detection encounters significant challenges due to scarce and imbalanced sample availability, where rigorous manufacturing processes yield extremely limited defect samples with skewed distributions, compromising deep learning models' ability to effectively capture defect characteristics. To overcome these limitations, comprehensive solutions integrating data augmentation techniques, transfer learning from pre-trained models, generative adversarial networks (GANs) for realistic defect synthesis, and active learning strategies for intelligent sample annotation can be implemented. This multifaceted approach not only enhances dataset representativeness but also substantially improves detection accuracy under data-constrained conditions [15].

4.4. Insufficient National and Industry Standardization

The current chip defect detection technologies face standardization challenges due to the absence of unified defect definitions, inspection procedures, and evaluation criteria, making it difficult to compare and mutually recognize results across different institutions—a critical bottleneck for industrial adoption. To address this, it is imperative to establish an industry-wide standardization framework that clearly defines core specifications including defect classification, detection methodologies, accuracy metrics, and data formats. Concurrently, promoting industry-academia-research collaboration through shared databases and public testing platforms will facilitate technical exchange, accelerate the development of standardized inspection protocols, and ultimately provide reusable technical specifications and practical guidelines for the sector [16].

5. Conclusion

This paper comprehensively reviews the research status, key technologies, and application prospects of image-based chip defect detection. It examines the application of deep learning algorithms (e.g., YOLO, Faster R-CNN) in chip defect inspection, analyzes their strengths and limitations, and summarizes various improved algorithms designed to address challenges such as micro-defect detection and complex background interference. Furthermore, the study identifies four major challenges in chip defect detection: (1) the trade-off between high-precision detection and cost, (2) balancing detection speed with production efficiency, (3) scarcity and imbalanced distribution of defect samples, and (4) insufficient national and industry standardization—proposing technical solutions including transfer learning and generative adversarial networks (GANs).

Looking ahead, image-based chip defect detection will focus on technological innovation, enhanced automation and intelligence, and the democratization of AI platforms. Advanced AI algorithms and multi-modal fusion technologies will significantly improve detection accuracy and efficiency, while fully automated inspection systems with real-time feedback mechanisms will become mainstream, enabling end-to-end automation from image acquisition to defect identification and reducing reliance on manual labor. Moreover, specialized AI detection platforms will lower technical barriers,

empowering more enterprises to enter chip manufacturing and fostering industry-wide adoption and development. These trends will collectively drive the semiconductor manufacturing industry toward greater intelligence and operational efficiency.

References

[1] Y. He, M. Xin, Y. Wang and C. Xu, Automatic Edge Detection Method of Power Chip Packaging Defect Image Based on Improved Canny Algorithm, 2024 9th International Conference on Electronic Technology and Information Science (ICETIS), Hangzhou, China, pp. 281-286, 2024.

[2] H. Ren, M. Tian, G. Zhang and W. Zhou, A Multi-Scale Feature Fusion Network for Chip Surface Defect Detection, 2024 IEEE International Conference on Image Processing (ICIP), Abu Dhabi, United Arab Emirates, pp. 957-962, 2024.

[3] H. Huang, Z. Shao, Q. Cheng and X. Wu, Multi-Scale Feature Fusion Network for Remote Sensing Image Captioning, 2023 30th International Conference on Geoinformatics, London, United Kingdom, pp. 1-4, 2023.

[4] Y. Cao, Y. Ni, Y. Zhou, H. Li, Z. Huang and E. Yao, An Auto Chip Package Surface Defect Detection Based on Deep Learning, in IEEE Transactions on Instrumentation and Measurement, vol. 73, pp. 1-15, 2024, Art no. 3507115,

[5] J. Zhao, W. Hu, D. Li, et al. COD-YOLO: An Efficient YOLO-Based Detector for Laser Chip Catastrophic Optical Damage Defect Detection. Arabian Journal for Science and Engineering, (prepublish):1-16, 2024.

[6] L. Abualigah, A. Diabat, S. Mirjalili, et al., The arithmetic optimization algorithm, Computer methods in applied mechanics and engineering, vol. 376, pp. 113609, 2021.

[7] K. Chen, H. Yao and Z. Han, Arithmetic optimization algorithm to optimize support vector machine for chip defect Identification, 2022 28th International Conference on Mechatronics and Machine Vision in Practice (M2VIP), Nanjing, China, pp. 1-5, 2022.

[8] J. Lu, et al., Research on Defect Recognition of Ceramic Chips for High Temperature Fuel Cells Based on Improved Faster R-CNN, 2022 International Conference on Machine Learning and Knowledge Engineering (MLKE), Guilin, China, pp. 143-150, 2022.

[9] S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks, in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 6, pp. 1137-1149, 1 June 2017.

[10] Shirui Technology. Information on: http://dlsense.com/. 2025/4/10

[11] Rongshi Intelligent Technology Co., Ltd, Information on: http://www.rovision.com.cn/?page_id=38. 2025/4/10

[12] Clear Soft Micro Vision, Information on: https://www.tvision-ai.com/about. 2025/4/10

[13] B.X. Chen, As 2/3nm Nodes Approach, Chip Testing Becomes Increasingly Complex. China Electronics News, 2022.

[14] W.Q. Liu, K. Chen, B. Wu, et al., Energy-Efficient and Secure Emerging Computing Chips: Current Status, Challenges, and Prospects. Scientia Sinica Informationis, 54(01): 34-47, 2024.

[15] Y.H. Li, Research on Chip Defect Detection Methods Based on Computer Vision. Hebei University of Science and Technology, 2025.

[16] K. Lou, Y.Y. Sun, H.H. Bai, Current Status and Advances in Automotive Chip Testing Standards. China Standardization, (19): 278-283, 2024.