Computers and Artificial Intelligence；Vol.3, No.3; 2026
E-ISSN: 3006 -3728 P-ISSN：3006-371X
DOI: https://doi.org/10.70267/cai.26v3n3.98103
Published by: Zeus Press

A Review of Deep Learning-Based 15Visual Inspection Techniques for Chip Surface Defects

Borui Cui*, Yuhuai Yin, Yujing Xiao, Sijia Wang, Xiao Xiao

School of Mathematics, Tianjin University, Tianjin 300350, China

**Corresponding author: Borui Cui.*

Abstract

With the continuous improvement of semiconductor chip integration, chip surface defects have a serious impact on chip performance and reliability. Traditional detection methods such as manual inspection and traditional machine vision have problems such as strong subjectivity, reliance on manual feature design, and poor adaptability, which are difficult to meet the high-precision detection needs of advanced chips. Deep learning has shown obvious advantages in automatic feature extraction, small target detection and complex background adaptation, and has gradually become the core technology of chip surface defect visual inspection. This paper systematically combs the types and imaging characteristics of chip surface defects, classifies and summarizes deep learning-based detection methods from four aspects: classification, object detection, segmentation and Transformer architecture, and analyzes the key technical difficulties such as multi-scale defects, complex texture interference and data imbalance in the detection process. On this basis, the current challenges in detection accuracy, cross-process generalization, data annotation and industrial deployment are discussed, and the future development trends such as multi-modal fusion, weakly supervised learning, model lightweight and intelligent quality control closed-loop are prospected. The review aims to provide theoretical reference and technical support for the research and engineering application of chip surface defect detection based on deep learning.

Keywords

chip surface defect, deep learning, visual inspection, defect detection, semiconductor manufacturing

1. Introduction

A semiconductor chip (i.e., an integrated circuit chip) is a circuit chip in which semiconductor components are encapsulated in a substrate such as metal, plastic, or ceramic to perform specific functions [1]. As integration levels continue to rise, surface defects such as scratches, pinholes, foreign particle contamination, and metal bridging severely impact the electrical performance and reliability of chips; therefore, efficient and accurate surface defect detection is of paramount importance. Among traditional methods, manual inspection is highly subjective and inefficient; while traditional machine vision is objective, it relies on manually designed features, has poor adaptability to complex or unknown defects, and struggles to meet the high-precision quality inspection requirements of advanced chips. In recent years, breakthroughs in deep learning within the field of computer vision have demonstrated its effectiveness in automatic feature extraction for inspections of painted

surfaces [2], solar cells [3], and steel plate surfaces [4]. Compared to traditional methods, deep learning can autonomously learn discriminative features from data. It excels particularly in detecting minute defects against complex textured backgrounds, eliminating the need for manually defined rules and significantly improving accuracy and adaptability [5]. Furthermore, it can achieve good performance with limited samples through transfer learning. This paper systematically reviews deep learning-based visual inspection methods for chip surface defects, analyzes key technical issues and engineering strategies, and discusses current challenges and future trends to provide a reference for related research and engineering practice.

2. Types of Surface Defects on Chips and Fundamentals of Imaging

2.1 Classification of Typical Defects and Image Features

During chip manufacturing and packaging and testing, factors such as process fluctuations, equipment aging, and environmental contamination can introduce various surface defects. Deep learning-based visual inspection relies on models to autonomously learn the distribution of image features associated with defects; therefore, establishing a physical classification of defects and identifying their characteristics—such as size, shape, and contrast—in optical images is a prerequisite for developing efficient detection algorithms.

Based on their physical form and formation mechanisms, typical defects on chip surfaces are classified into three categories: structural damage (such as scratches and chipped edges), contamination and foreign particles (such as particles and stains), and pattern and circuit defects (such as bridging, open circuits, and pinholes). These defects present three major challenges in automatic optical inspection (AOI) images: First, there is extreme multiscale complexity: macro-scale defects can span hundreds of pixels, while micro-scale defects may be only a few pixels in size, requiring the model to possess the ability to fuse multiscale features (such as feature pyramids or multiscale feature enhancement modules) [6]. Second, random shapes and diverse orientations—circuit defects depend on circuit geometry and topology, while scratches and stains have irregular shapes—necessitating the use of deformable convolutions or attention mechanisms to extract complex shape features. Third, strong interference from contrast and texture: low-contrast defects blend with the background, complex wiring textures obscure minute defects, and highly reflective metal surfaces are prone to strong artifacts, whose gradients may exceed those of actual defects and trigger false positives. The above feature analysis provides a critical basis for designing deep learning inspection algorithms that integrate multi-scale fusion, geometric adaptability, and anti-interference capabilities.

2.2 Common Imaging Methods

High-quality image acquisition is a physical prerequisite for deep learning inspection algorithms. Due to the highly reflective nature of chip surfaces and their complex microstructures, customized optical solutions are required, primarily including brightfield/darkfield illumination, line-scan cameras, and high-magnification microscopy.

Bright-field illumination uses a coaxial or high-angle light source to capture specular reflections; a smooth substrate appears as a bright background, while defects appear as dark areas, making it suitable for detecting macroscopic damage. Dark-field illumination uses a low-angle light source; light reflected from a smooth surface appears as a dark background, while fine particles and microcracks produce diffuse scattering and appear as bright targets. This method is sensitive to height variations and is suitable for detecting sub-pixel-level defects. High-end equipment often employs bright-field and dark-field stroboscopic composite imaging, where dual-source images are acquired and fused via multi-channel network fusion [7]. Through this fusion mechanism, complementary features that are easily lost under single-illumination conditions can be effectively integrated, significantly enhancing the contrast and detection rate of minute defects. For large-area chips in continuous motion, line-scan cameras synchronize the horizontal scan frequency with the motion speed via an encoder, stitching each line into a seamless, high-resolution image to eliminate motion blur and stitching artifacts; linear light sources ensure grayscale consistency; TDI technology improves the signal-to-noise ratio at high speeds. For sub-micron defects, high-magnification microscopy employs high-numerical-aperture objectives and short-wavelength light sources to overcome the diffraction limit. At high magnifications, the depth of field is extremely shallow, requiring the combination of autofocus and focus stacking to generate fully sharp images for analysis by CNNs or ViTs.

2.3 Characteristics of Chip Image Data

Images of chip surfaces exhibit three key characteristics: complex background textures, minute defects, and class imbalance. Complex background textures include high-frequency patterns formed by high-density interconnects, periodic lattices in memory areas, and irregular interlacing in logic areas; defect edges are easily obscured, leading to high false positive rates. Additionally, minute gray-scale fluctuations caused by processes such as CMP are visually similar to actual defects, requiring models to possess high-dimensional semantic understanding. Defects are minute, with critical defects typically spanning only 3 × 3 to 5 × 5 pixels. After multiple downsampling steps in the main network, features of small defects are easily overwhelmed by the background, leading to missed detections; extremely small targets are sensitive to bounding box shifts, causing a sharp drop in IoU, and making it difficult to allocate high-quality positive samples during training. Class imbalance: With wafer yields exceeding 99%, normal samples constitute the overwhelming majority, while defects are rare. The loss function is dominated by normal samples, causing the model to tend to classify everything as normal. Defects also exhibit a long-tail distribution internally, with high-frequency defects like scratches at the head and rare defects at the tail—or even as small samples—leading to overfitting at the head and poor generalization. Furthermore, chip data is considered trade secrets, public datasets are scarce, and annotation costs are high, all of which exacerbate the imbalance. To address these typical industrial challenges of small sample sizes and long-tail distributions, the use of Generative Adversarial Networks (GANs) combined with transfer learning for data augmentation has become a key technical approach for effectively expanding training sets for rare defects and mitigating class imbalance [8].

3. Deep Learning-Based Defect Detection Methods

Currently, defect detection on chip surfaces still relies primarily on manual visual inspection and traditional machine vision, but both approaches have significant drawbacks. Manual visual inspection is prone to operator fatigue, has high rates of missed and false detections, and is characterized by high labor costs and slow processing speeds. Traditional machine vision relies on manually designed features (such as edge operators and texture filters) in conjunction with classifiers; however, it suffers from poor generalization capabilities when dealing with complex, minute, and variable defects, and often fails when lighting conditions or chip models change. Deep learning-based defect detection methods can automatically learn hierarchical features from data to achieve end-to-end mapping [9], offering three key advantages: strong adaptability (requiring only data updates and retraining), high accuracy (10–30 percentage points higher than traditional methods), and end-to-end integration. Consequently, nearly all recent research on chip surface defect detection has relied on deep learning. Deep learning methods can be categorized into those based on classification, object detection, segmentation, and Transformer architectures.

3.1 Classification-Based Methods

Classification methods determine the defect category of image patches. In practice, an image patch classification strategy is employed: the large image is divided into small patches (e.g., 128 × 128), and a CNN (such as VGG or ResNet) is used to determine the presence of defects in each patch, which are then stitched together to form a defect heatmap. ResNet addresses the degradation issue in deep networks through residual modules; ResNet50 achieves an accuracy of 90.23% in chip classification tasks [10]. This classification method features a simple structure, high speed, and excellent performance for defects with single, well-defined rules, while transfer learning reduces data requirements. Its limitations include the inability to obtain precise defect contours, a tendency to miss defects that span multiple blocks, and reduced recognition accuracy for imbalanced classes.

3.2 Object Detection-Based Methods

Object detection simultaneously outputs the location (bounding box) and class of a defect. Two-stage detectors (such as Faster R-CNN) first generate region proposals and then perform fine-grained classification; while they offer high accuracy, they are slow. Single-stage detectors (such as YOLO) perform the entire process in a single step; while they are fast, they are insensitive to small objects. To address minute defects, feature pyramids (FPN) are used to fuse multi-level features, and attention mechanisms (such as the SE module)

are employed to enhance key regions. The improved YOLO-ER framework effectively enhances the detection accuracy of minute objects [11].

3.3 Segmentation-Based Methods

Segmentation methods enable pixel-level classification and accurately delineate defect contours. The U-Net architecture, a representative model, utilizes skip connections to preserve fine details; DeepLab employs dilated convolutions and spatial pyramid pooling to capture multi-scale context [12]. Segmentation methods are suitable for the precise delineation of irregular defects (such as burrs and forked cracks) and the quantitative analysis of defect geometric parameters (length, area), thereby meeting the requirements for precision inspection of high-reliability chips.

3.4 Transformer-Based Methods

Transformers, based on self-attention mechanisms, can directly establish global dependencies. Self-attention enables the network to compare global information, making it more sensitive to minor defects. The representative model DETR treats detection as a set prediction, resulting in a streamlined process; Swin Transformer reduces computational complexity through sliding windows and outperforms CNNs on images with complex textures and low-contrast chip patterns. For example, the unsupervised ViT method achieved an AUROC of 0.951 on the MVTec-3D dataset [13].

4. Conclusion

As integrated circuit processes advance toward 3nm and smaller nodes, defects on chip surfaces are becoming increasingly minute, diverse, and concealed. Existing deep learning-based detection technologies face systemic challenges in terms of detection limits, generalization capabilities, data dependency, and industrial deployment. Detecting microscopic defects (<10 nm, sub-pixel scale) requires joint optimization of multi-modal fusion at the imaging end (brightfield, darkfield, SEM) and multi-scale features combined with attention mechanisms at the algorithmic end; super-resolution reconstruction can further enhance resolution capabilities. Insufficient cross-process generalization—deploying a TSMC N3 model to Samsung's 3GAA process caused the mAP to drop from 95% to below 70%. Domain adaptation and meta-learning (fine-tuning with 5–10 samples) can improve cross-process adaptability [14], but building a universal dataset still requires industry collaboration. Annotation costs are exorbitant, with pixel-level annotation of a single high-resolution image taking over 30 minutes, and rare defects accounting for less than 1%; Self-supervised learning (MAE, restoring 92% performance with 10% of the labeled data) [15] and generative models (StyleGAN3) for synthesizing defects offer viable solutions to alleviate the data shortage. Industrial deployment faces stringent real-time requirements (full inspection of a 12-inch wafer within 30 minutes), and high-precision models (Swin Transformer, DeepLabv3+) consume significant resources; Model lightweighting (YOLOv8 INT8 quantization achieves 4x speedup with <1% accuracy loss) [16] and dedicated hardware acceleration (FPGA and Huawei Ascend 310 achieving 100 frames per second throughput) have become practical solutions; in the future, a unified interface will be needed to achieve closed-loop quality control.

Overall, deep learning technology has fundamentally transformed the landscape of chip defect detection, with some advanced production lines achieving accuracy rates exceeding 99%. However, challenges remain, including insufficient interpretability, room for improvement in robustness, and limitations to surface defect detection; Future efforts should focus on developing explainable AI, multimodal fusion (surface → internal defects), weakly supervised/zero-shot detection [14], and deep integration with smart manufacturing to build an intelligent quality control system spanning the entire design-manufacturing-inspection chain, thereby providing core support for the high-quality development of China's semiconductor industry.

References

[1] Du Zhongyi. (2012). *Semiconductor Chip Manufacturing Technology*. Electronics Industry Press: Beijing.

[2] Dai Liang, Ju Yifan, Fu Zhengqiang, et al. Research on Deep Learning-Based Methods for Detecting Surface Defects in Coated Surfaces [J/OL]. Modern Coatings and Coating Technology, 1–13 [2026-05-08]

[3] Wang Xianbao, Li Jie, Yao Minghai, et al. (2014). A Deep Learning-Based Method for Solar Cell Surface Defect Detection. *Pattern Recognition and Artificial Intelligence*, *27*(06)

[4] He Di. (2021). *Application of Deep Learning in Steel Plate Surface Defect and Character Recognition*. University of Science and Technology Beijing.

[5] Wang Yikun. (2022). *Research on Machine Vision-Based Algorithms for Flat Chip Surface Defect Detection*. Heilongjiang: Harbin Engineering University.

[6] Wang Xiang, Huang Juan, Gu Jinan, et al. (2026). Model for Chip Surface Defect Detection Based on Directional Decoupling and Multiscale Enhancement. *Semiconductor Technology*, *51*(1), 68–76.

[7] Ye Qixin. (2021). *A Study on Panel Microdefect Detection Methods Based on Bright-Field and Dark-Field Fusion*. Harbin: Harbin Institute of Technology.

[8] Cao Pengbin, Lei Zhenglong, Chen Yanbin, et al. (2022). A Study on Defect Detection of Laser Soft-Soldered Joints with Few Samples Based on GAN Data Augmentation and Transfer Learning. *China Laser*, *49*(16), 1602014.

[9] Chen Miao et al. AI-Empowered Textile Composites: Applications, Challenges, and the Future of Deep Learning. *Journal of Composite Materials*, 1–16.

[10] Xia Wensheng et al. (2025). A Study on High-Resolution Remote Sensing Image Classification Based on ResNet50. *Science and Technology Innovation and Application*, *15*(35), 72–74+79.

[11] Qiu Jiawei et al. YOLO-ER: An Efficient Multi-Scale Object Detection Framework for Complex Environments. Research on Computer Applications: pp. 1–10.

[12] Su Nan and Zhao Zhenhua. A Semantic Segmentation Method for UAV Images Based on an Improved DeepLab V3+ Network. *Journal of the Naval Aviation University*, 1–11.

[13] Wang Junmin, Fu Jingfei, and Ning Chaokui. (2026). Industrial Defect Detection Based on Unsupervised Learning and Visual Transformers. *Journal of Datong University*, *Shanxi (Natural Science Edition)*, *42*(01), 115–120.

[14] H. Li, X. Li, Q. Fan, Q. Xiong, X. Wang, and V. C. M. Leung. (2024). Transfer Learning for Real-Time Surface Defect Detection With Multi-Access Edge-Cloud Computing Networks. *IEEE Transactions on Network and Service Management*, *21*(1), 310–323.

[15] M. Caron et al. (2023). DINOv2: Learning robust visual features without supervision. arXiv:2304.07193, 2023.

[16] B. Jacob et al. (2018). Quantization and training of neural networks for efficient integer-arithmetic-only inference. in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR).

Funding

This research received no external funding.

Conflicts of Interest

The authors declare no conflict of interest.

Acknowledgment
This paper is an output of the science project.

Copyrights
Copyright for this article is retained by the author (s), with first publication rights granted to the journal. This is an open - access article distributed under the terms and conditions of the Creative Commons Attribution license (http://creativecommons. org/licenses/by/4. 0/).