Dean&Francis

ISSN 2959-6149

A Comprehensive Investigation on Convolutional Neural Network-Based Product Defect Detection in Industrial Applications

Shaoqin Chen

College of Information Science and Electronic Engineering, Zhejiang University, Zhejiang, China
3220103761@zju.edu.cn

Abstract:

Ensuring high product quality has become increasingly important in modern manufacturing. Traditional manual inspection and basic machine vision systems often perform not well in accuracy, efficiency, and robustness, especially in complex industrial environments. Convolutional Neural Networks (CNNs), which is a technology in deep learning specialized for visual data, have shown great potential in product defect detection. A broad range of recent CNN-based methods for product defect detection is systematically reviewed in this article by classifying existing methods into three major types: baseline CNN models, attention-based models, and hybrid CNN frameworks. Baseline models such as MobileNet, ResNet, and EfficientNet focus on improving accuracy and computational efficiency. Attention-based models incorporate spatial and channel attention mechanisms to localize subtle defects better. Hybrid models combine CNNs with techniques like Random Forests or adaptive fusion strategies to improve robustness across varied defect types and product conditions. In addition to summarizing technical advancements, this paper highlights three critical challenges that hinder real-world deployment: (1) lack of model interpretability, (2) limited applicability, and (3) difficulty in deploying models on resource-constrained edge devices. To address these, this paper further discusses promising solving directions, including the integration of domain knowledge and expert systems, domain adaptation and generalization strategies, and model optimization methods like model pruning, reduced-precision quantization, and student–teacher distillation frameworks. Overall, this review provides a systematic overview, offering valuable insights for researchers and engineers, and explores future research paths for enhancing CNN-based defect detection in manufacturing.

Keywords: Convolutional neural network; defect detection; industrial inspection.

ISSN 2959-6149

1. Introduction

In modern manufacturing industries, product quality is a key factor in customer satisfaction, market competitiveness, and enterprise reputation. A defect in a product can not only influence its function and safety, but also lead to tremendous loss on economy, product recalls, and serious brand damage. As industries move toward intelligent and automated production lines, ensuring high product quality through accurate and timely defect detection has become a critical challenge.

Traditionally, product defect detection majorly relies on manual inspection or simple machine vision systems [1]. Although manual inspection is intuitive, it is often labor-consuming, uncontinuous, and tend to cause manul error, especially under high throughput conditions. Simple machine vision system methods, on the other hand, lack flexibility and robustness when faced with complex or subtle defect patterns. These limitations have created a growing demand for intelligent, scalable, and high-accuracy inspection solutions.

Recent advancements in Artificial Intelligence (AI), especially in deep learning techniques, have significantly brought numerous new possibilities to product defect detection. For example, CNNs, a classical model of deep learning specifically designed for visual data, have demonstrated exceptional performance in a variety of computer vision tasks ranging from assigning class labels to images, identifying object locations, to understanding pixel-level image content [2]. Because they can automatically extract multi-level visual representations from raw images, CNNs are particularly tailored for detecting subtle and complex product defect detection, where defects often appear in various shapes, sizes, and textures etc.

A growing number of researches have explored approaches based on CNNs for product defect detection across various industrial domains. These approaches have demonstrated strong performance in tasks of product defect detection such as surface inspection, assembly line quality control, and component integrity assessment. For instance, in 2017, Wang et al. developed a fast and robust CNN model, which was particularly designed for detecting defects in product quality inspection, achieving high accuracy and efficiency in real-world industrial manufacture [3]. More recently, combining with deep CNNs and machine vision, a product defect detection architecture was proposed by Kaushal A Desai et al. in 2022, further improving the robustness and generalization of defect detection under varying conditions [4]. Additionally, the availability of standardized datasets such as DAGM, MVTec [5], and KolektorSDD [6], has provided a solid foundation for CNNs-based methods in product defect detection domain.

Despite the rapid progress in CNN-based defect detection, several critical challenges still exist. One major issue is the lack of interpretability which means CNNs are often regarded as black boxes. This is especially a serious problem in manufacturing environments with high risks, where incorrect predictions can lead to significant financial losses. Another concern is the limited applicability of models. Many CNNs perform well only within narrowly defined domains or under specific conditions or even on specified datasets. When applied to different product types, materials, or production settings, their performance often becomes worse, requiring frequent retraining and data collection. Furthermore, deployment on edge devices is difficult to realize due to limited memory and computational resources. Real-time tasks, compatibility with industrial software, and long-term maintainability all present practical difficulties. All these issues highlight the need for a systematic review that not only summarizes the state-of-the-art methods, but also critically appoints key limitations such as interpretability, adaptability, and deployment constraints.

Motivated by the importance of product quality assurance and the rapid growth of CNNs-based solutions, this paper aims to provide a comprehensive survey of recent advances in CNNs-based product defect detection. The objective is to summarize and analyze the current methods, highlight their contributions and differences, and point out challenges to be solved. The structure of the paper is arranged as follows. In Section 2, an overview of some CNNs-based defect classification methods is presented. Section 3 provides a critical discussion of current limitations, emerging trends and future research. In the end, Section 4 concludes the total article.

2. Methods

In the domain of product defect classification, due to excellent capacity to learn hierarchical features from visual data, CNNs have shown remarkable success. Section 2 outlines the major categories of CNN-based methods which are applied to product defect detection, including baseline CNN architectures, attention-enhanced CNN models, and hybrid CNN models.

2.1 Baseline CNN Models

Baseline CNN models serve as the foundational backbone for defect classification tasks. These networks, initially developed for image classification, have been widely adopted in industrial applications owing to their robustness and adaptability.

2.1.1 MobileNet

MobileNet, firstly proposed by Howard et al., is a lightweight CNNs architecture designed for efficient computation on mobile devices [7]. It makes use of separable convolutions to reduce the number of parameters and computational cost, making it suitable for real-time defect classification on the resource-constrained devices.

In product quality inspection tasks, Zhang et al. proposed an improved MobileNetV2-SSDLite architecture which was tailored for fabric defect detection under a cloud-edge collaborative framework. As a lightweight adaptation of the Single Shot MultiBox Detector (SSD), SSDLite reduced computational complexity by applying depthwise separable convolutions in the detection module. The model was deployed at the edge devices for real-time fabric defect detection, while the cloud platform was responsible for data management and training, realizing a real-time and high-efficiency fabric defect detection pipeline [8].

2.1.2 ResNet

In the PCB manufacturing industry, detecting cosmetic defects is becoming increasingly challenging due to the complexity of modern boards. To solve this problem, Zhang et al. proposed CS-ResNet, a cost-sensitive variant of ResNet-50 tailored for PCB inspection. The model maintains ResNet's residual architecture with shortcut connections to ease gradient flow and stabilize training. A key innovation is the cost-sensitive adjustment layer added after the fully connected layer. This layer reweights the loss based on class distribution and misclassification cost, allowing the model to focus more on rare real defects and reduce false alarms. The overall pipeline includes standard CNN components: convolution, batch normalization, ReLU, ResBlocks, global average pooling, and a final cost-sensitive output. This design improves accuracy and recall on imbalanced datasets while maintaining high efficiency, making CS-ResNet well-suited for real-time PCB quality control systems [9].

2.1.3 EfficientNet

EfficientNet, proposed by Tan and Le, is a family of CNNs. Through jointly adjusting the model's number of layer, channels per layer, and input resolution, it can enhance accuracy and efficiency [10]. Compared to traditional CNNs, EfficientNet significantly reduces the number of parameters and FLOPs while maintaining the same accuracy or even improving, making it suitable for deployment in real-world industrial inspection tasks.

Several recent studies have shown that EfficientNet offers notable improvements in handling surface defects in product inspection. For example, one research integrated EfficientNet with YOLOv5 for steel surface defect detection, demonstrating notable improvements in mean average precision and classification accuracy [11]. In this approach, the original YOLOv5 backbone network is replaced with EfficientNet variants, aiming to leverage their superior feature extraction capabilities. Another study evaluated EfficientNet-B3 against MobileNetV2 and ResNet-50 on a steel defect classification task, and the results showed that EfficientNet-B3 achieved a higher recall rate with fewer false positives, indicating its superior robustness in identifying subtle defects [12].

2.2 Attention-Based CNN Models

Attention mechanisms enhance CNNs by allowing them to focus on the most relevant regions of an input image, which is particularly useful in defect detection where defects are often subtle.

2.2.1 Complementary Attention Network (CAN)

To improve defect detection accuracy under complex backgrounds, Zhao et al. proposed the CAN, which was equipped with an advanced dual-branch attention module [13]. The key innovation lies in combining channel-wise attention and spatial attention to allow the network to focus on both "what" and "where" to look when identifying defects in solar cell electroluminescence images.

In the channel-wise attention subnetwork, Global Average Pooling (GAP) and global max pooling (GMP) features are concatenated and passed through convolution operations. This design preserves more discriminative information. Next, the spatial attention subnetwork processes the channel-refined features to generate a spatial attention map that emphasizes relevant spatial locations.

Overall, CAN acts as a plug-and-play attention filter that enhances CNN representational power, especially under complex background disturbances. Experiments demonstrate that CAN significantly outperforms traditional attention mechanisms such as CBAM in solar cell defect detection, making it a promising solution for industrial quality inspection tasks.

2.2.2 Object-Level attention mechanism with Bi-CAM

Hu and Wang proposed an efficient CNN model tailored for detecting defects in complex and various industrial scenarios [14]. A novel object-level attention mechanism is integrated into the model, enabling it to localize and emphasize regions containing defects during training. This is achieved without the need for additional network structures. In order to improve the model's capacity to distinguish subtle differences, they incorporate bilinear pooling layers, allowing interactions between feature channels.

Furthermore, to improve interpretability of the model and better suit the bilinear architecture, the authors propose a variant of traditional CAM named Bilinear Class Activation Maps (Bi-CAM). Bi-CAM serves as a visualization technique to help to highlight the specific image regions that contribute to the model's predictions.

ISSN 2959-6149

2.3 Hybrid CNN Models

Hybrid models combine CNNs with other machine learning algorithms to enhance complementary strengths and improve robustness, accuracy and other performance parameters in varying datasets or conditions.

2.3.1 Convolutional Neural Network and Random Forest (CNN-RF)

In 2024, Banerjee et al. proposed a hybrid CNN-RF model for accurate casting defect forecasting [15]. The integration of deep learning techniques with conventional machine learning methods has shown notable effectiveness in defect classification tasks. In this approach for defect detection, a CNN is first used to extract high-level visual features from input images. Instead of relying on a fully connected layer for classification, these features are passed to a Random Forest (RF) classifier.

This two-stage framework takes advantage of CNNs' ability to learn discriminative representations and RF's robustness, especially when dealing with small datasets or datasets with noise. RF is then trained on CNN-extracted features, enhancing overall accuracy and interpretability. Results demonstrate improved detection performance compared to using CNN alone.

2.3.2 Hybrid attention network with ASFF

Aiming to solve the problems in detecting intricate or fine-grained steel surface defects, Zhou et al. proposed a hybrid attention-based convolutional architecture that enhances detection accuracy through three key innovations [16].

Firstly, to improve feature representation by combining both spatial and channel attention mechanisms, the Convolutional Block Attention Module (CBAM) was introduced. Moreover, a module named Adaptively Spatial Feature Fusion (ASFF) was applied to the neck of the network to fuse information across different feature layers. This allows the model to assign optimal weights to features from varying resolutions, significantly improving the detection of defects with various sizes and shapes. Lastly, the Complete IoU (CIOU) loss function was applied to improve bounding box regression accuracy and enhance generalization. Built on YOLOX-S, this hybrid attention network demonstrated significant performance gains on detecting fine-grained defects in the NEU-DET benchmark dataset, outperforming standard models such as SSD, YOLOv3. These enhancements demonstrate the effectiveness of integrating multiple attention mechanisms and adaptive fusion strategies into CNN-based architectures for industrial defect detection tasks.

3. Discussion

Despite recent advances in CNN-based defect detection models, the field still faces several crucial challenges that limit the practical deployment and scalability of these systems. These challenges are primarily rooted in the inherent limitations of deep learning models and their interaction with complex real-world and real-time industrial scenarios. This paper discusses three main challenges—interpretability, applicability, and deployment constraint and then outlines promising directions which have the possibility to address these challenges.

3.1 Challenges and Limitations

3.1.1 Interpretability

A persistent challenge in deploying CNNs for industrial defect detection tasks is poor interpretability of models. As being called black-box models, CNNs often produce predictions without providing humans with convincing reasons. This opacity is particularly a serious problem in manufacturing environments which aims to high quality of products. For instance, classifying a non-defective product as defective product by mistake could result in unnecessary waste and financial loss, while missing a defect may lead to safety hazards or brand damage. Moreover, some defects such as micro cracks, subtle surface abrasions are inherently difficult to localize or classify because of their small sizes or irregular patterns. Current interpretability techniques, like CAM or Bi-CAM, provide some insight into this problem. However, these methods often lack consistency and do not align well with engineering judgment. This gap between model output and domain reasoning remains a crucial issue to be solved.

3.1.2 Limited applicability

Another major concern is the poor generalization ability of many CNN-based models. While high performance can be achieved on specific datasets or well-controlled product lines, model accuracy usually decreases when model is applied to different product types, materials, or manufacturing environments. For example, some slight modifications to a product's design such as slight changes in texture, shape, or surface, can result in inefficacy of previously training on the model. This lack of robustness means that companies must frequently collect new datasets and then retrain models on these datasets to adapt to each product variation, resulting in increased costs and delayed deployment. Given the dynamic nature of industrial manufacturing in the real world, the inability of models to adapt to novel conditions remains a bottleneck in defect detection tasks.

3.1.3 System deployment

Beyond model performance problems, a major challenge of CNN-based defect detection tasks lies in the deployment and maintenance of these models in industrial systems. Unlike training under specific datasets, factory environments are highly compositive, often composed of different systems such as control modules, production pipelines and other complicated platforms. Thus, deploying a CNN model in such a setting requires compatibility with real-time communication protocols, integration with factory management software, and seamless coordination with sensors on the production pipelines.

And in many cases, deployment is hindered by limited computational resources on edge devices, or the lack of standardization in data formats. Even after deployment, models may require regular updates due to evolving defect patterns or changes in production processes. However, frequent retraining and revalidation are rarely feasible in practice due to the need for production line stability and the high cost of time. Furthermore, industrial users often lack access to specialized AI teams, making the long-term maintenance, debugging, and adaptation of deep learning models particularly challenging. Without specialized tools for monitoring model, updating parameters, and managing failure cases, the system's reliability can degrade over time.

3.2 Future Prospects

To address the above challenges, several promising research directions and technical innovations have emerged. While each of these approaches is still evolving, they offer ideas to improve model interpretability, adaptability, and other bottleneck problems.

3.2.1 Incorporating domain knowledge and expert systems

To improve interpretability and reduce reliance on datasets, integrating expert knowledge into CNN models presents a convincing strategy. Expert systems, rule-based constraints can help CNNs make decisions that align with human reasoning. For example, incorporating defect features or domain knowledge into the loss function or feature design could enhance model credibility and robustness. Similarly, attention mechanisms informed by engineering rules can guide the model to focus on physically relevant regions. This hybrid approach mixes the learning capacity of CNNs with the interpretability.

3.2.2 Domain adaptation and domain generalization

Since CNNs often perform not well beyond training data, improving their adaptability to diverse real-world environments is essential. Two important research directions have gained increasing attention: domain adaptation and domain generalization.

Domain adaptation focuses on helping a model trained on one type of product or dataset work well on another type, even if the two differ in appearance, materials, or other conditions. In particular, unsupervised domain adaptation allows the model to adjust to a new environment without many labeled samples. Techniques like adversarial learning [17], aligning feature distributions have shown good results [18].

On the other hand, domain generalization aims to train models that can naturally perform well on unseen domains, without any fine-tuning. This is useful when new products or defect types are introduced. This method helps the model focus on general patterns rather than domain-specific details, making it more reliable across different conditions.

3.2.3 Lightweight and efficient architectures

In real-world manufacturing environments, many defect detection systems need to run directly on edge devices such as cameras, sensors, or embedded boards rather than powerful servers with abundant resources. These devices have limited memory and processing ability, so CNN models deployed on them require lightweight and efficient architectures. To meet the real-time and efficiency requirements under such limitations, some researches have been conducted on model optimalization techniques that transform standard deep neural networks into lightweight and deployable architectures.

Model pruning is a widely adopted method which means cutting out the parts of a neural network that aren't really needed. By removing these unnecessary weights and connections, the architectures of models become smaller and have a higher speed, while using less memory and computing resources. Moreover, a technique called quantization has the ability to reduce the size of the numbers or arrays used in the model. This method also can make the model much lighter and helps it run more quickly on edge devices with limited resources. Besides, there is another method called knowledge distillation that transfers knowledge from a high-capacity teacher model to a smaller student model, effectively retaining accuracy while substantially reducing model size and complexity. The student picks up the key knowledge from the teacher and becomes almost as accurate as the teacher, and meanwhile has a much smaller size and faster speed which is perfect for deployment on devices with limited capacity. In all, these techniques can be applied individually or jointly to compress large-scale models without substantial loss of performance.

ISSN 2959-6149

4. Conclusion

This paper presents a comprehensive review of recent developments in CNNs for product defect detection in industrial settings. By analyzing representative approaches across baseline CNN architectures, attention-based models, and hybrid frameworks, this article summarizes the major trends and strategies in this field.

The methods mentioned above demonstrate strong potential in accurately identifying various types of surface defects in product. Lightweight models offer promising solutions for real-time applications, while attention mechanisms and hybrid models improve detection precision in complex scenarios. Through discussion, this article identifies several key challenges that limit current methods, including poor interpretability, limited applicability across product domains, and practical deployment difficulties in factory architectures. These issues highlight the gap between academic research and industrial needs.

To solve these bottleneck problems, future research should explore integration of domain knowledge, expert systems, and advanced techniques to enhance model robustness and adaptability. Overall, CNN-based defect detection is a rapidly evolving area with significant promise for improving quality control in intelligent manufacturing.

References

[1] Xie X. A review of recent advances in surface defect detection using texture analysis techniques. ELCVIA: Electronic Letters on Computer Vision and Image Analysis, 2008: 1–22.

[2] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature, 2015, 521(7553): 436–444.

[3] Wang T, Chen Y, Qiao M, Snoussi H. A fast and robust convolutional neural network-based defect detection model in product quality control. The International Journal of Advanced Manufacturing Technology, 2018, 94: 3465–3471.

[4] Singh SA, Desai KA. Automated surface defect detection framework using machine vision and convolutional neural networks. Journal of Intelligent Manufacturing, 2023, 34(4): 1995–2011.

[5] Bergmann P, Fauser M, Sattlegger D, Steger C. MVTec AD—A comprehensive real-world dataset for unsupervised anomaly detection. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019: 9592–9600.

[6] Tabernik D, Šela S, Skvarč J, Skočaj D. Segmentation-based deep-learning approach for surface-defect detection. Journal of Intelligent Manufacturing, 2020, 31(3): 759–776.

[7] Howard AG, Zhu M, Chen B, Kalenichenko D, Wang W, Weyand T, et al. Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv Preprint, 2017, arXiv:1704.04861.

[8] Zhang J, Jing J, Lu P, Song S. Improved MobileNetV2-SSDLite for automatic fabric defect detection system based on cloud-edge computing. Measurement, 2022, 201: 111665.

[9] Zhang H, Jiang L, Li C. CS-ResNet: Cost-sensitive residual convolutional neural network for PCB cosmetic defect detection. Expert Systems with Applications, 2021, 185: 115673.

[10] Tan M, Le Q. EfficientNet: Rethinking model scaling for convolutional neural networks. Proceedings of the International Conference on Machine Learning, 2019: 6105–6114.

[11] Ren F, Fei J, Li H, Doma BT Jr. Advancements in steel surface defect detection: An enhanced YOLOv5 algorithm with EfficientNet integration. International Journal of Electrical and Electronics Research, 2024, 12(1): 268–275.

[12] Khan IU, Aslam N, Aboulnour M, Bashamakh A, Alghool F, Alsuwayan N, et al. Deep learning-based surface defect detection in steel products using convolutional neural networks. Mathematical Modelling of Engineering Problems, 2024, 11(11).

[13] Su B, Chen H, Chen P, Bian G, Liu K, Liu W. Deep learning-based solar-cell manufacturing defect detection with complementary attention network. IEEE Transactions on Industrial Informatics, 2020, 17(6): 4084–4095.

[14] Hu C, Wang Y. An efficient convolutional neural network model based on object-level attention mechanism for casting defect detection on radiography images. IEEE Transactions on Industrial Electronics, 2020, 67(12): 10922–10930.

[15] Banerjee D, Sharma N, Upadhyay D, Singh V. Hybrid CNN-RF model for accurate casting defect forecasting. Proceedings of the 2024 Asia Pacific Conference on Innovation in Technology (APCIT), 2024: 1–6.

[16] Zhou M, Lu W, Xia J, Wang Y. Defect detection in steel using a hybrid attention network. Sensors, 2023, 23(15): 6982.

[17] Awais M, Zhou F, Xu H, Hong L, Luo P, Bae SH, Li Z. Adversarial robustness for unsupervised domain adaptation. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021: 8568–8577.

[18] Zhang C, Zhang J. Transferable regularization and normalization: Towards transferable feature learning for unsupervised domain adaptation. Information Sciences, 2022, 609: 595–604.