Paradigm Evolution of Industrial Surface Defect Detection: The Underlying Logic and Fundamental Challenges from Supervised Classification to Unsupervised Anomaly Localization

Zihan Zhang

Beijing University of Technology - Dublin International College, Beijing University of Technology, Beijing, China
zihan.zhang@ucdconnect.ie

Abstract. Industrial appearance defect detection is a key aspect of quality control in smart manufacturing. The core challenge lies in how to effectively apply models that perform well in laboratory environments to complex, dynamic, and unpredictable real-world industrial scenarios. This challenge is specifically manifested in the sharp contradiction between the infinite possibilities of defect types in industrial production lines and the extreme scarcity of labeled data. This paper systematically reviews the paradigm evolution in this field, from closed-set fully supervised classification to open-set anomaly detection, and further to open-vocabulary semantic understanding. First, an in-depth analysis was conducted on the inherent limitations of fully supervised models based on U-Net and ResNet regarding label dependency and distribution shift issues. Secondly, it explores how unsupervised paradigms (such as PatchCore and PaDiM) achieve unknown defect detection by learning only normal samples, as well as their shortcomings in distinguishing defect types. Finally, it describes how the open-vocabulary paradigm based on vision-language large models (such as CLIP) provides new approaches for zero-shot defect classification through semantic guidance. By comparing the performance and underlying logic of various paradigms on benchmark datasets such as MVTec AD and VisA, this paper reveals the development trends in the field of industrial visual inspection and provides theoretical guidance and practical directions for building the next generation of highly adaptable and robust industrial detection systems.

Keywords: Industrial appearance defect detection; Fully supervised learning; Unsupervised anomaly localization; Open-vocabulary detection

1 Introduction

Industrial appearance defect detection is a core component of smart manufacturing and automated quality control systems, and its performance directly affects product yield, production efficiency, and corporate competitiveness. With the in-depth advancement of national strategies such as 'Industry 4.0' and 'Made in China 2025,' achieving

© The Author(s) 2026
K. Subramanian (ed.), *Proceedings of the International Workshop on Advances in Deep Learning for Image Analysis and Computer Vision (IWADIC 2025)*, Advances in Computer Science Research 128,
https://doi.org/10.2991/978-94-6239-648-7_75

intelligent and digitalized production processes has become an inevitable trend in the upgrading of the manufacturing industry. In this context, machine vision-based automated defect detection technology is of crucial practical significance for enhancing the core competitiveness of the manufacturing industry, reducing labor costs, and achieving closed-loop quality control.

However, this field has long faced a fundamental challenge: "Models that are well-trained in simple and controllable laboratory environments find it difficult to maintain stable performance in complex, dynamic, and unknown real-world industrial environments"[1]. The essence of this issue lies in the fact that the types of defects in real industrial production are virtually limitless, and obtaining a large number of accurately labeled defect samples is extremely costly, if not infeasible. Therefore, researching intelligent detection models that can adapt to open environments and do not rely on extensive defect labeling is not only a hot topic at the academic frontier but also key to advancing industrial vision from 'laboratory demonstration' to 'production line implementation'.

In order to systematically outline the technological evolution path in this field, this paper focuses on several representative research works. The U-Net proposed by Ronneberger et al. pioneered the use of an encoder-decoder structure with skip connections, achieving breakthroughs in biomedical image segmentation. Its architecture has become a benchmark model for subsequent industrial defect pixel-level segmentation tasks, laying the foundation for the fully supervised paradigm [2]. PatchCore, proposed by Roth et al., serves as a performance benchmark for unsupervised anomaly localization paradigms. This method constructs a feature memory bank of normal samples and performs efficient nearest neighbor search, achieving near-perfect detection accuracy on the MVTec AD dataset, demonstrating the powerful potential of the 'learning only from normal' approach [3]. WinCLIP, proposed by Jeong and others, represents the latest open-vocabulary paradigm. This method is optimized for the CLIP model and achieves zero-shot defect classification and localization through regional contrast and prompt integration, providing a pioneering solution to the 'semantic gap' problem of unsupervised methods [4]. In addition, the MVTec LOCO dataset released by Bergmann et al. systematically differentiates between structural anomalies and logical anomalies for the first time, promoting the evolution of evaluation standards from merely detecting visual differences to higher-level semantic and functional understanding, and posing new challenges for the next generation of defect detection methods [5].

The above content clearly outlines the evolutionary trajectory of industrial defect detection from closed-set to open-set, and then to open-vocabulary paradigms. This paper aims to systematically review this process, deeply analyze its intrinsic logic and fundamental challenges, and provide insights for future research directions.

2 The Paradigm Evolution of Industrial Defect Detection

2.1 Paradigm One: Fully Supervised Learning

Fully supervised learning is the fundamental paradigm for industrial defect detection, and its core assumption is that the testing environment is consistent with the training environment, and all defect categories that need to be detected appear in the training set. This paradigm learns a mapping function from input images to defect categories or pixel-level segmentation maps through end-to-end training.

In terms of technical implementation, two architectures played a key role: ResNet solves the problems of gradient vanishing and degradation in very deep networks through residual learning, providing powerful feature extraction capabilities for high-precision image-level defect classification [6]. U-Net, with its encoder-decoder structure and skip connections, achieves precise pixel-level defect segmentation and has become the standard solution for defect localization [2].

However, this paradigm has fundamental limitations. First, it cannot detect any defect types that did not appear in the training set, and this 'closed-world' assumption is seriously inconsistent with the openness of real industrial environments. Secondly, obtaining a large amount of pixel-level annotated data requires extremely high expert and time costs. Research shows that on the MVTec AD dataset, producing sufficient pixel-level annotations for just a single category requires about 40-80 person-hours [1]. These limitations constitute the direct impetus for the emergence of new paradigms.

2.2 Paradigm Two: Unsupervised Anomaly Localization

The unsupervised anomaly localization paradigm proposes a fundamental shift: using only normal samples for training, and detecting any anomaly that does not conform to the modeled distribution of "normal". This shift inherently enables the detection of unknown defects and completely eliminates the reliance on annotated defect data [7].

In terms of methodological evolution, PaDiM establishes a multivariate Gaussian distribution for each image location and generates anomaly maps by calculating the Mahalanobis distance between test features and these distributions. It achieved 97.5% pixel-level AUROC on MVTec AD while maintaining high efficiency [7]. PatchCore [3], by constructing a memory bank of normal features and using coreset sampling, achieved a record-breaking 99.6% image-level AUROC on MVTec AD, reducing the image-level classification error rate by more than half compared to previous state-of-the-art methods [3].

A comparative summary of the performance and key characteristics of these representative methods across different paradigms is provided in Table 1. The advantages of the unsupervised paradigm are evident: complete independence from defect annotations and the ability to detect unknown defects. However, its core limitation lies in the inability to distinguish specific defect types (e.g., scratch, dent), i.e., the "semantic gap". Additionally, models are sensitive to new normal patterns not covered in the training set (e.g., new lighting, angles), making them prone to false

alarms. These limitations prompted the development of new paradigms capable of semantic understanding.

Table 1. Performance Comparison of Representative Methods of Each Paradigm on the MVTec AD Dataset

paradigm	Representative method	Training Data Requirements	Image-level AUROC	Pixel-level AUROC	Key advantage
Fully supervised	U-Net	Defect Normal (pixel-level annotation)	~90%	~85%	Accurate in locating known defects
Unsupervised	PaDiM	Only normal samples	Not applicable	97.5%	Balancing efficiency and accuracy
Unsupervised	PatchCore	Only normal samples	99.6%	~98%	Unsupervised SOTA accuracy
Open vocabulary	WinCLIP	Zero-shot (no training required)	91.8%	~85%	Zero-shot, distinguishable defect types
Open vocabulary	MemSeg	Pretrained model for normal samples only	95.2%	96.3%	Combining the advantages of large models and memory databases

2.3 Paradigm Three: Open Vocabulary / Large Model Driven

The open-vocabulary paradigm utilizes prior knowledge learned from massive internet data by vision-language foundation models (e.g., CLIP), defining and detecting defects through natural language prompts. This achieves a paradigm shift from "teaching the model to recognize defects using data" to "telling the model what defect to look for using language".

Technically, the CLIP model established a general association between visual concepts and language descriptions through contrastive learning on 400 million image-text pairs, enabling zero-shot recognition [8]. Building on this, WinCLIP achieved pixel-level zero-shot anomaly classification and segmentation via compositional prompt ensemble and window-based feature extraction and aggregation, reaching 91.8% image-level AUROC on MVTec AD [4].

Recent research further explores optimizing large models for industrial scenarios. MemSeg combines a pre-trained vision-language model with a memory bank mechanism, enhancing normal feature representation and employing anomaly suppression strategies. It achieved 95.2% image-level AUROC on MVTec AD, demonstrating the potential of combining unsupervised ideas with the advantages of large models [9].

The performance of these open-vocabulary and related few-shot methods on MVTec AD and VisA datasets is quantitatively compared in Table 2. The value of this paradigm lies in its unprecedented flexibility – users can dynamically define new defects by modifying text prompts without retraining the model, preliminarily bridging the "semantic gap". However, as an emerging paradigm, its localization accuracy generally still lags behind dedicated unsupervised methods, and it faces challenges such as "hallucination" risk and high computational cost.

Table 2. Comparison of Image-level AUROC (%) Between Zero-Shot and Few-Shot Methods on MVTec AD and VisA

Methods	Settings	MVTec AD	VisA
CLIP-AC	0-shot	74.0	59.3
WinCLIP	0-shot	91.8	78.1
SPADE	1-shot	81.0	79.5
PatchCore	1-shot	83.4	79.9
WinCLIP+	1-shot	93.1	83.8

3 Standard Datasets and Evaluation Criteria

3.1 Guidance Dataset

To fairly compare various paradigms, this paper introduces the following public datasets:

MVTec AD is the most authoritative benchmark in the field, containing 5,354 high-resolution images across 15 categories (5 textures and 10 objects), covering 73 types of defects [1]. Its provision of high-quality pixel-level ground truth makes it the primary standard for evaluating the generality of methods.

VisA contains 12 categories and more than 10,000 images. Its multi-instance object characteristics increase the complexity of localization, making it more suitable for evaluating the robustness of models in complex scenarios [10].

MVTec LOCO innovatively introduced the distinction between 'logical anomalies' (such as incorrect or missing parts) and 'structural anomalies' (such as scratches or dents) [5]. It is specifically designed to assess a model's ability to understand the functional and semantic relationships of scenes, posing a direct challenge to the emerging open-vocabulary paradigm.

3.2 Inspection Indicators

This article adopts a multi-level evaluation system to comprehensively measure model performance: Image-level AUROC is used to assess the model's overall ability to distinguish between normal and abnormal images, suitable for defect presence detection tasks. Pixel-level AUROC evaluates the model's accuracy in locating abnormal regions and is the key metric for defect segmentation tasks. To further eliminate the interference of threshold selection strategies. The F1-max Score reflects the model's discriminative ability purely by traversing all possible decision thresholds and taking the maximum F1 score. In addition, PRO/sPRO fairly evaluate the model's ability to detect tiny defects by treating each defective area and even each pixel equally, making them the strictest segmentation evaluation standards currently available.

3.3 Shortcomings of Datasets and Evaluation Metrics

The Closedness and Idealization of the Dataset. Images in current mainstream benchmarks (such as MVTec AD) are usually well-calibrated and aligned, which differs from the complexity, multi-view, and lighting variations found in real industrial environments. Although studies such as Rd-MVTec AD have begun to focus on this issue and existing benchmark datasets are helpful, there is a need to establish more non-aligned datasets of different scales and modalities.

Evaluation criteria fail to fully reflect the application value. The evaluation criteria fail to fully reflect the practical value. Metrics such as pixel-level AUROC are widely used, but they may not completely represent the core needs of industrial clients. For example, it may place too much emphasis on detecting major defects while lacking sensitivity to the omission of minor but important defects. Future evaluations should focus more on recall rate, or metrics directly associated with the consequences of decisions made by automated production lines.

4 Performance and Analysis of Mainstream Methods

To quantitatively compare the performance across paradigms, this section provides a comprehensive analysis based on the results of representative methods summarized in Table 1 (in Section 2.2).

As clearly shown in Table 1, unsupervised methods represented by PatchCore have surpassed fully supervised methods on MVTec AD, achieving an image-level AUROC of 99.6%. This indicates that by training solely on normal samples, the model can near-perfectly determine whether an image contains an anomaly. This breakthrough not only validates the effectiveness of the "identify anomalies by learning normal samples" technical path but also makes it the mainstream choice in current industrial inspection. Meanwhile, open-vocabulary paradigms such as WinCLIP achieved 91.8% image-level AUROC under zero-shot conditions, demonstrating the feasibility of this class of

methods. MemSeg further shows that by combining pre-trained foundation models with unsupervised mechanisms, performance can be enhanced while maintaining semantic understanding capabilities. However, these methods still lag behind PatchCore (approx. 98%) in pixel-level AUROC (approx. 85% to 96.3%), reflecting that their localization accuracy is not yet on par with specialized models. Their core value lies not in higher numerical values but in a functional disruption—being able to preliminarily answer "what is this defect?".

Further testing on datasets like MVTec LOCO, which contains logical anomalies, reveals deeper issues: unsupervised methods excel at handling structural anomalies but their performance significantly drops on logical anomalies due to the lack of semantic understanding of object function and assembly relationships. In contrast, open-vocabulary methods demonstrate a relative advantage in logical anomaly detection by leveraging their inherent semantic capabilities. This indicates that when facing the complex challenges of an open world, relying solely on low-level visual feature contrast is insufficient, and semantic understanding ability is becoming increasingly critical.

5 Existing Problems and Future Directions

5.1 Fundamental Challenges of Each Paradigm

The fully supervised paradigm is still limited by annotation costs and the closed-world assumption, making it less applicable in industrial scenarios that require rapid adaptation to new products [1]. Although the unsupervised paradigm addresses the problem of detecting unknown defects, it has issues such as a 'semantic gap' and sensitivity to new normal patterns, resulting in a higher false alarm rate during production line changeovers. As an emerging direction, the open-vocabulary paradigm still needs improvement in terms of positioning accuracy, hallucination risk, and computational cost, and it is still some distance away from industrial-grade applications [4].

5.2 Future Research Directions

Based on the above analysis, this paper believes that future research should focus on the following directions. Firstly, in terms of paradigm integration, a hybrid architecture can be explored that combines the precise localization capability of unsupervised models with the semantic understanding capability of open-vocabulary models. For example, MemSeg has demonstrated the potential of such fusion, and subsequent research could further explore more innovative structural designs such as dual-branch networks [9]. Secondly, in terms of efficiency optimization, it is necessary to use techniques such as knowledge distillation and model pruning to compress the semantic understanding capabilities of large models into a computational budget that is acceptable in industrial environments. To achieve millisecond-level inference speed and meet real-time detection requirements. In addition, regarding the adaptive mechanism, online learning and incremental memory mechanisms are introduced. It

enables the model to quickly adapt to newly emerging normal patterns on the production line without a complete reconstruction, thereby effectively reducing the false alarm rate caused by production line changes. Finally, in terms of improving the evaluation system, it is necessary to go beyond the current single evaluation model dominated by AUROC. Develop an evaluation framework that comprehensively considers detection accuracy, inference speed, system robustness, and actual business value to more fully reflect the real application requirements in industrial scenarios.

6 Conclusion

This article systematically reviews the evolution of industrial appearance defect detection from fully supervised to unsupervised, and then to open-vocabulary paradigms, drawing the following main conclusions:

First of all, the main line of paradigm evolution is clear and distinct: The annotation requirements decrease at each level, the defect coverage expands at each level, and adaptability increases at each level. The performance from U-Net to PatchCore and then to WinCLIP clearly demonstrates the technical feasibility of "detecting defects without learning them", as well as the great potential for achieving semantic understanding under zero-shot conditions.

Secondly, there are structural obstacles behind the high targets. There is a gap between the idealized conditions of existing benchmark datasets and the complexity of real industrial environments, resulting in a decline in model performance in cross-device and cross-product generalization experiments. Scenarios requiring semantic understanding, such as logical anomalies, remain blind spots for unsupervised methods, while open-vocabulary methods still need to improve in localization accuracy.

Finally, future research should focus on the balance between 'accuracy and flexibility'. Through technical approaches such as paradigm integration, efficiency optimization, and adaptive mechanisms, developing a new generation of industrial vision systems that combine the localization accuracy of unsupervised methods with the semantic flexibility of open-vocabulary approaches. The ultimate goal is to ideally achieve the inclusion of new defects 'zero-sample' detection within a ten-minute deployment cycle.

References

1. Bergmann, P., Fauser, M., Sattlegger, D., et al.: MVTec AD — A Comprehensive Real-World Dataset for Unsupervised Anomaly Detection//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 9592-9600. (2019)
2. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation//International Conference on Medical Image Computing and Computer-Assisted Intervention. Cham: Springer, 234-241. (2015)
3. Roth, K., Pemula, L., Zepeda, J., et al.: Towards total recall in industrial anomaly detection//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 14318-14328. (2022)

4. Jeong, J., Zou, Y., Yeo, S., et al.: WinCLIP: zero-shot anomaly detection via region-based discrimination//Proceedings of the IEEE/CVF International Conference on Computer Vision. 1955-1964. (2023)
5. Bergmann, P., Batzner, K., Fauser, M., et al.: The MVTec Logical Constraints (LOCO) Dataset for Detecting Logical Anomalies//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 1-10. (2023)
6. He, K., Zhang, X., Ren, S., et al.: Deep residual learning for image recognition//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 770-778. (2016)
7. Defard, T., Levieux, G., Romet, A.: PaDiM: a patch distribution modeling framework for anomaly detection and localization//IEEE International Conference on Image Processing. IEEE, 475-479. (2021)
8. Radford, A., Kim, J. W., Hallacy, C., et al.: Learning transferable visual models from natural language supervision//International Conference on Machine Learning. PMLR, 8748-8763. (2021)
9. Wang, X., Wang, J., Li, L., et al.: MemSeg: A semi-supervised method for image surface defect detection using differences and commonalities. Engineering Applications of Artificial Intelligence, 123: 106457. (2023)
10. Zou, Y., Jeong, J., Pemula, L., et al.: Spot the difference: A novel task for embodied AI in changing environments//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 19102-19112. (2022)

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (http://creativecommons.org/licenses/by-nc/4.0/), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.