Self-Supervised Learning for Industrial Visual Anomaly Detection: A Review of Recent Advances, Applications, and Open Challenges

Author: Bao Tang

Abstract

Industrial visual anomaly detection (VAD)—the automated identification and localization of defects, irregularities, and deviations in manufactured products—plays a critical role in ensuring product quality, operational safety, and process reliability across modern manufacturing. The inherent scarcity of labeled anomaly data, the diversity of defect types, and the requirement for real-time deployment pose fundamental challenges that traditional supervised learning approaches struggle to address. In response, **self-supervised learning (SSL)** has emerged as a transformative paradigm, enabling models to learn rich representations from abundant unlabeled normal data by defining pretext tasks that do not require manual annotations. This review provides a comprehensive and critical synthesis of recent advances in self-supervised learning for industrial visual anomaly detection. We examine the methodological landscape across five major SSL categories—contrastive learning, masked reconstruction, generative modeling, rotation prediction, and cross-modal pretext tasks—and map their application to key industrial domains including surface inspection, 3D component quality control, semiconductor fabrication, and predictive maintenance. A structured analysis of eight representative works—including the Iterative Mask Reconstruction Network (IMRNet), graph attention-based multivariate anomaly detection, and diffusion-enabled defect synthesis—grounds the discussion in empirical evidence. We further explore the integration of SSL with digital twin platforms, the role of foundation models, and the unique challenges of real-world deployment. Finally, we identify open research problems and articulate a forward-looking agenda for the field.

Keywords: Self-Supervised Learning; Industrial Visual Anomaly Detection; Defect Inspection; Masked Reconstruction; Diffusion Models; Digital Twins; Predictive Maintenance; Smart Manufacturing

1. Introduction

The relentless drive toward zero-defect manufacturing—mandated by stringent quality standards in automotive, aerospace, electronics, and pharmaceutical industries—has placed unprecedented demands on automated visual inspection systems. Industrial visual anomaly detection (VAD) aims to identify product defects that deviate from nominal specifications, ranging from surface scratches and dents to structural voids and dimensional inaccuracies. Conventionally, this task relied on human inspectors or rule-based machine vision systems; however, the escalating complexity of manufactured products, the speed of production lines, and the need for sub-millimeter detection precision have rendered these approaches insufficient (Zhang et al., 2025; Liu et al., 2024).

The last decade has witnessed a paradigm shift driven by deep learning. Supervised convolutional neural networks (CNNs) trained on large labeled datasets of defective and non-defective samples have achieved human-competitive detection accuracy in well-defined industrial settings (Liu et al., 2024). However, supervised approaches face a fundamental obstacle in industrial practice: **anomaly data is intrinsically scarce**. Defective samples constitute only a tiny fraction of production output—often less than 1%—and manually collecting, annotating, and maintaining representative defect datasets is expensive, time-consuming, and prone to bias. Furthermore, the space of possible defects is open-ended: new failure modes emerge as materials, processes, and product designs evolve. Supervised models trained on a finite set of known defect categories inevitably fail to generalize to novel anomalies—precisely the cases where detection matters most.

These limitations have catalyzed intense interest in **self-supervised learning (SSL)** for industrial VAD. SSL enables models to learn meaningful representations from large volumes of unlabeled data by solving pretext tasks derived from the data itself. In the industrial context, the core intuition is elegant: if a model can learn to represent the structure, geometry, and appearance of **normal** products accurately, then deviations from this learned normalcy can be flagged as anomalies—without ever needing explicit examples of the defects themselves. This makes SSL particularly well-suited to the industrial setting, where normal data is abundant and labeled anomalous data is scarce.

This review synthesizes the rapidly growing body of literature on SSL for industrial VAD. Our contributions are threefold:

1. **Methodological taxonomy**: We propose a structured classification of SSL approaches for industrial VAD, identifying five major paradigms and their representative algorithms.
2. **Application mapping**: We systematically map SSL methods to key industrial domains—surface inspection, 3D metrology, semiconductor fabrication, and predictive maintenance—highlighting empirical advances and deployment considerations.
3. **Research agenda**: We identify open challenges, emerging trends (including diffusion models, digital twin integration, and foundation models), and promising future directions.

The review is organized as follows: Section 2 provides background on industrial VAD; Section 3 introduces the SSL paradigm; Section 4 presents the methodological taxonomy; Section 5 maps methods to applications; Section 6 discusses cross-cutting themes; and Section 7 offers a forward-looking conclusion.

---

2. Background: Industrial Visual Anomaly Detection

2.1 Problem Formulation and Industrial Context

Industrial visual anomaly detection encompasses a family of tasks with varying levels of granularity: **anomaly detection** (is the sample defective or not?), **anomaly localization** (where is the defect?), and **anomaly classification** (what type of defect is it?). The open-set nature of the anomaly class—encompassing all possible deviations from normal—fundamentally distinguishes VAD from standard supervised classification, where all classes are predefined during training.

The industrial imperatives for robust VAD are compelling. In automotive manufacturing, undetected surface defects on painted body panels result in costly rework or customer complaints. In semiconductor fabrication, micro-cracks or particle contaminants on wafers can cause complete chip failure. In pharmaceutical production, pill defects such as chipping or discoloration are regulatory compliance issues. Across these domains, the economic

consequences of missed defects are substantial, while false positive detections impose unnecessary production interruptions and waste (Liu et al., 2024).

2.2 Limitations of Traditional Approaches

Classical machine vision approaches to VAD relied on hand-crafted feature extractors (e.g., Sobel edges, HOG descriptors, Gabor filters) combined with statistical classifiers or rule-based decision thresholds. While interpretable and computationally efficient, these methods suffer from limited representational capacity and poor generalization to novel defect types. They require substantial domain-specific engineering and are brittle under variations in illumination, pose, or material appearance.

The transition to deep learning promised a data-driven solution, but early supervised CNNs replicated the fundamental bottleneck: they require large, balanced, and accurately annotated datasets of both normal and anomalous samples. In industrial practice, this requirement is rarely satisfiable. Anomaly datasets are not only small but also inherently biased—presenting only the defect types that have been observed and documented—which creates a dangerous illusion of comprehensive coverage (Zhang et al., 2025).

Beyond purely visual inspection, industrial quality assurance increasingly leverages multi-physics sensing to capture structural and thermal signatures that are invisible to conventional cameras. Huang and colleagues (2023) developed a four-dimensional thermal imaging system for non-uniform surfaces, demonstrating that the integration of structured illumination binocular cameras with infrared thermography enables accurate reconstruction of temperature fields across complex geometries—an approach that has direct implications for thermally-triggered anomaly detection in manufacturing. Similarly, Huang and colleagues (2026) combined stereo phase-measuring deflectometry with deep learning-enhanced phase unwrapping to achieve precision 3D surface metrology of optical components, illustrating how learned representations can augment physics-based sensing for higher-fidelity defect detection.

2.3 The Open-Set Challenge and the Case for Self-Supervised Learning

The open-set nature of industrial anomalies demands a fundamentally different learning paradigm. Rather than learning to discriminate among a fixed set of defect classes, the goal is to learn the manifold of normal appearance—and to flag samples that deviate significantly from this manifold as anomalous. This is precisely the objective of **anomaly detection by reconstruction**: if a model trained exclusively on normal data cannot faithfully reconstruct an input, the reconstruction error serves as an anomaly score (Zhang et al., 2025).

Self-supervised learning operationalizes this principle by designing pretext tasks that teach the model the structure of normal data without labels. By solving tasks such as "reconstruct the masked regions of this image" or "predict the rotation angle of this object," the model is forced to learn rich, hierarchical representations of normal appearance—representations that can then be used for zero-shot anomaly detection (Liu et al., 2024).

---

3. Self-Supervised Learning: Principles and Paradigms

3.1 Core Principle

Self-supervised learning derives supervisory signals from the structure of the data itself, bypassing the need for manual labels. The training signal is computed from a **pretext task**—an auxiliary task designed to require semantic understanding of the data—and a **downstream task**—the ultimate target application. In industrial VAD, the downstream task is almost universally anomaly detection or localization, while the pretext task defines how the model learns from normal data.

A critical property of SSL in the industrial context is that it trains on **normal data only**. This aligns perfectly with the reality of manufacturing: production lines generate abundant normal samples, while anomalous samples are rare, costly, and potentially dangerous to collect at scale. By learning what normal looks like, SSL sidesteps the data imbalance problem that undermines supervised approaches.

3.2 Five Paradigms for Industrial VAD

The SSL methods applied to industrial VAD can be organized into five major paradigms, each exploiting a different structural property of normal data:

Contrastive Learning (SimCLR, MoCo, BYOL): Learn representations by pulling augmented views of the same image closer together (positive pairs) while pushing views of different images apart (negative pairs). In the industrial VAD context, the assumption is that normal samples share structural similarity under augmentation, while anomalous samples exhibit reconstruction patterns that break this similarity (Liu et al., 2024).

Masked Reconstruction (MAE, IMRNet): Train a model to reconstruct randomly masked patches of an input image. The model must develop a holistic understanding of object structure, texture, and geometry to fill in missing regions. Anomalies—characterized by unusual textures or missing structural elements—are poorly reconstructed, yielding high reconstruction error (Li et al., 2024; Zhang et al., 2025).

Generative Modeling (VAE, GAN, Diffusion): Train a generative model on normal data to learn the probability distribution of normal samples. Anomalies are detected as low-probability samples under this distribution. Recent advances in diffusion models have enabled high-fidelity synthesis of normal appearance, improving the quality of the learned manifold and reducing false positive rates (Zhang et al., 2025; Khan et al., 2025).

Rotation Prediction (RotNet): Train a model to predict the rotation angle (0°, 90°, 180°, 270°) at which an input image was rotated. The model must learn to recognize object semantics and structure to solve this task—knowledge that transfers to anomaly detection, as anomalies disrupt the rotational symmetry and canonical orientation of normal objects (Liu et al., 2024).

Cross-Modal Pretext Tasks: Leverage complementary modalities (e.g., 3D point clouds, depth maps, infrared images) to define pretext tasks that require cross-modal understanding. Anomalies often manifest differently across modalities, and multi-modal SSL can exploit these differences for more robust detection (Li et al., 2024).

---

4. Methodological Landscape

4.1 Masked Reconstruction: IMRNet and Beyond

Masked reconstruction has emerged as one of the most influential SSL paradigms for industrial VAD. The core methodology—masking random patches of an input and training a model to reconstruct them—forces the network to develop a comprehensive understanding of global structure and local texture. Reconstruction error serves as a natural anomaly score: normal regions are accurately reconstructed, while anomalous regions produce high error.

Li and colleagues (2024) introduced the **Iterative Mask Reconstruction Network (IMRNet)** for 3D anomaly detection, addressing one of the most pressing challenges in industrial quality control: the scarcity of annotated 3D anomaly data. Point cloud data from 3D sensors captures geometric information that is largely invariant to surface texture and lighting conditions, but 3D anomaly datasets are substantially smaller than their 2D counterparts. IMRNet tackles this by combining a geometry-aware point-cloud sampling (GPS) module with a self-supervised masked reconstruction strategy. During training, the network learns to reconstruct randomly masked point patches, developing a structural model of normal 3D geometry. At inference, regions that deviate significantly from the learned normal structure yield high reconstruction errors and are flagged as anomalies (Li et al., 2024).

The design of IMRNet reflects a broader principle in SSL for industrial inspection: **domain-specific pretext tasks outperform generic ones**. Rather than applying a standard masked autoencoder directly to point clouds, IMRNet incorporates geometry-aware sampling that respects the local structural regularity of manufactured components—patches from flat or smoothly curved surfaces are more predictable than those near sharp edges or complex geometry, allowing the model to learn a nuanced model of structural normality (Li et al., 2024).

Complementing IMRNet, a recent study by Zhang and colleagues (2025) proposed a masked reconstruction framework for multivariate industrial time-series anomaly detection, combining graph attention mechanisms to extract sequence correlation features with a variational autoencoder (VAE) encoding-decoding network. This work illustrates the versatility of the masked reconstruction principle across data modalities, showing that the core idea—learn to reconstruct normal patterns, detect deviations—transfers effectively from vision to multivariate sensing data.

4.2 Contrastive Learning and Anomaly Detection by Similarity

Contrastive learning has been extensively explored for industrial VAD through methods such as SimCLR-based and BYOL-based frameworks. The central idea is to learn an embedding space where normal samples from the same class are close to each other while being distant from samples of different classes. A notable application in industrial settings is the use of **memory banks** to store representative embeddings of normal product variants, enabling fast nearest-neighbor-based anomaly scoring at inference.

Research by Liu and colleagues (2024) surveyed deep learning approaches for industrial image anomaly detection, documenting multiple contrastive learning instantiations that achieve competitive results on benchmark datasets such as MVTec AD. A key finding is that data augmentation strategies—such as color jitter, geometric transforms, and CutMix—play a critical role in determining the quality of learned representations. Augmentations that preserve the semantic identity of normal samples while introducing controlled variations enable the model to learn invariant and robust features.

4.3 Generative Models: VAEs, GANs, and Diffusion

Generative approaches have a long history in anomaly detection, rooted in the intuition that a model trained on normal data will assign low likelihood to anomalous inputs. Early work used Variational Autoencoders (VAEs) and Generative Adversarial Networks (GANs); recent advances have shifted toward **diffusion probabilistic models** (diffusion models), which have demonstrated superior fidelity in modeling complex data distributions.

Zhang and colleagues (2025) highlighted the emerging role of diffusion models in industrial defect detection and data augmentation. Diffusion models learn to synthesize data by denoising random noise through an iterative refinement process, enabling the generation of high-fidelity normal samples that can augment training datasets for downstream supervised models. This is particularly valuable in industrial settings where certain defect categories may be entirely absent from historical datasets—a fundamental limitation that purely discriminative approaches cannot overcome (Zhang et al., 2025).

A study on few-shot steel surface defect generation (Khan et al., 2025) demonstrated that diffusion models conditioned on defect-free reference images can generate realistic synthetic defect samples with minimal training data. This capability addresses one of the most persistent challenges in industrial VAD: the data imbalance between defect classes and the prohibitive cost of collecting rare defect examples. By generating diverse synthetic defects, diffusion models enable more comprehensive evaluation of detection systems and can even be used to fine-tune supervised models for rare defect categories (Khan et al., 2025).

The integration of latent diffusion models with defect segmentation networks has shown particular promise in steel surface inspection, where spatial resolution and fine-grained texture detail are critical. Stable diffusion models fine-tuned on normal surface textures produce high-quality reconstructions; anomalies are detected through pixel-level reconstruction error between the input and its reconstruction (Zhang et al., 2025).

4.4 Graph Attention-Based Multivariate Anomaly Detection

A particularly relevant recent contribution is the work by Zhang and colleagues (2025) on multivariate industrial time-series anomaly detection using **graph attention mechanisms combined with self-supervised variational autoencoders**. This approach addresses a critical gap in conventional univariate anomaly detection methods, which ignore the inter-variable dependencies that characterize real industrial processes—where temperature, pressure, vibration, and electrical signals are inherently coupled.

The proposed model operates in two stages. First, a **graph attention network** extracts sequence correlation features from multi-modal time-series data, adaptively weighting the influence of each variable based on learned attention coefficients. These features are concatenated with the original data to form a dual-feature representation. Second, this representation is fed into a **VAE encoding-decoding network** trained in a self-supervised manner on normal operational data. Anomalies are detected as inputs that the VAE cannot faithfully reconstruct—indicating that they deviate from the learned normal operational regime (Zhang et al., 2025).

This work exemplifies the broader trend of combining **structural modeling** (via graph neural networks) with **self-supervised representation learning** (via masked reconstruction or VAE), achieving robust anomaly detection in complex industrial processes where multiple sensors interact.

4.5 SSL for Digital Twin-Based Predictive Maintenance

A parallel and increasingly important application of SSL is in **predictive maintenance (PdM)** enabled by **digital twin (DT)** platforms. A digital twin is a high-fidelity virtual replica of a physical asset—machine, production line, or factory—that evolves in real time as sensor data streams in. By continuously comparing the observed behavior of the physical system with its digital counterpart, digital twins can detect anomalies that presage equipment failure, enabling proactive maintenance before breakdowns occur.

Khan and colleagues (2025) proposed a comprehensive **data-driven digital twin framework for predictive maintenance of smart manufacturing systems**, integrating machine learning models that learn normal equipment behavior from historical sensor data. The framework leverages multi-sensor fusion—combining vibration, acoustic, thermal, and electrical signals—to build a holistic model of equipment health. Anomalies are detected as deviations between the predicted state (from the digital twin model) and the observed state, with the magnitude of deviation correlating with the severity of the impending fault (Khan et al., 2025).

The connection to SSL is direct: SSL methods enable the digital twin model to learn robust representations of normal equipment behavior without requiring labeled examples of failure events—which, by definition, are rare and potentially catastrophic. Techniques such as masked reconstruction and contrastive learning applied to multivariate sensor time series can learn invariant representations of normal operating states, enabling zero-shot detection of novel failure modes (Khan et al., 2025).

Recent work on generative and predictive AI for digital twin systems in manufacturing (Frontiers in AI, 2025) further elaborated the integration of foundation models and large-scale pre-trained transformers with digital twin platforms. Pre-trained transformer models, initially developed for natural language processing, are being fine-tuned on domain-specific industrial data to enable more accurate fault detection, predictive maintenance scheduling, and real-time anomaly flagging. The combination of generative AI with digital twins also enables fault simulation—generating synthetic scenarios of equipment degradation to stress-test PdM algorithms and train maintenance personnel (Frontiers in AI, 2025).

4.6 Comparative Analysis

The following table summarizes the key characteristics of the SSL paradigms examined in this review:

Paradigm	Core Pretext Task	Data Modality	Key Strength	Key Limitation
Contrastive Learning	Positive/negative pair discrimination	2D images, time series	Learns discriminative embeddings	Sensitive to augmentation choice
Masked Reconstruction	Masked patch reconstruction	2D images, 3D point clouds, time series	Learns rich structural representations	Computational cost for high-resolution images
Generative (Diffusion)	Score matching / denoising	2D images, 3D point clouds	Highest fidelity normal data modeling	Slow inference; requires careful likelihood calibration
Rotation Prediction	Rotation angle classification	2D images	Simple, interpretable pretext	Limited to image data
Cross-Modal	Cross-modal prediction	Multi-modal (2D+3D, vision+language)	Exploits complementary modalities	Requires aligned multi-modal data

5. Application Domains

5.1 Surface Inspection

Surface inspection is the most established application domain for industrial VAD, encompassing tasks such as detecting scratches, cracks, stains, and dimensional deviations on flat or slightly curved product surfaces. The MVTec Anomaly Detection (MVTec AD) benchmark—a large-scale dataset of normal and defective industrial images—has catalyzed rapid progress, enabling fair comparison across methods. SSL approaches, particularly masked reconstruction (e.g., DRAEM, RIAD) and contrastive learning methods, have consistently outperformed supervised baselines on this benchmark, especially in the few-shot and zero-shot settings that mirror real industrial deployment conditions (Liu et al., 2024).

5.2 3D Component Quality Control

With the proliferation of 3D sensors (structured light scanners, LiDAR, fringe projection systems), 3D VAD has gained prominence as a complementary modality to 2D vision. IMRNet (Li et al., 2024) represents a landmark contribution in this domain, demonstrating that self-supervised masked reconstruction on point cloud data can achieve state-of-the-art performance on 3D anomaly benchmarks while requiring only normal training data. The 3D approach is particularly valuable for safety-critical components—such as turbine blades, medical implants, and structural aerospace parts—where internal defects are not visible in 2D images.

5.3 Semiconductor Fabrication

Semiconductor manufacturing imposes some of the most demanding VAD requirements: defects at the nanometer scale, inspection under严格的环境控制, and zero tolerance for yield loss. The application of SSL in semiconductor wafer inspection is an emerging area, where diffusion-based data augmentation has shown particular promise in generating realistic synthetic defect patterns

for training data augmentation (Khan et al., 2025).

5.4 Predictive Maintenance

Predictive maintenance extends VAD beyond product inspection to process and equipment health monitoring. Digital twin platforms (Khan et al., 2025; Wang et al., 2025) now integrate SSL-based anomaly detection with physics-based equipment models, enabling the early detection of incipient failures—such as bearing wear, motor winding degradation, and coolant system blockages—through analysis of multivariate sensor streams. Beyond passive monitoring, the integration of intelligent robotics into manufacturing lines introduces new inspection modalities: Li and colleagues (2024) demonstrated that collaborative robotic manipulators equipped with vision-based gesture interfaces can perform adaptive, real-time inspection tasks in dynamic environments, combining the flexibility of human-robot collaboration with automated anomaly detection capabilities. The convergence of robotic inspection, multi-sensor data fusion, and SSL-based anomaly detection points toward fully autonomous quality assurance systems capable of continuous learning and adaptation.

The causal reasoning capabilities underlying predictive maintenance are also benefiting from advances in structured neural architectures. Zhu and Liu (2026) showed that hybrid graph attention networks combined with LSTM enable causal-aware demand forecasting in supply chains—a capability that translates directly to equipment health prediction, where causal dependencies among sensor signals, operating conditions, and failure modes are explicitly modeled. Furthermore, Zhu and Liu (2025) integrated causal discovery with Bayesian graph neural networks for transparent environmental risk forecasting, demonstrating how causal graph structures learned from observational data can improve the interpretability and reliability of anomaly predictions in safety-critical systems.

---

6. Discussion

6.1 The Data Scarcity Paradox and Synthetic Data Generation

A fundamental tension in industrial VAD is the **data scarcity paradox**: the methods that require the most data (deep learning) are precisely those for which data is most scarce (anomalies). The emergence of diffusion models as a practical tool for **synthetic defect generation** offers a compelling resolution. By learning to model the distribution of normal product appearance, diffusion models can be inverted or conditioned to generate plausible defect samples—surfacing from this distribution in the direction of anomalousness. This synthetic data can augment training sets for supervised models, improve evaluation coverage, and enable data-driven testing of VAD systems under controlled defect scenarios (Zhang et al., 2025; Khan et al., 2025).

6.2 Foundation Models and Transfer Learning

The recent emergence of **vision foundation models** (e.g., CLIP, SAM, DINOv2) has opened a new frontier for industrial VAD. Pre-trained on internet-scale image datasets, these models encode rich semantic representations that transfer effectively to industrial inspection tasks. Self-supervised vision transformers such as DINO and DINOv2, trained via self-distillation on large uncurated datasets, learn visual features that exhibit remarkable robustness to texture variations, lighting changes, and geometric transformations—properties highly valuable in industrial settings (Zhang et al., 2025). Fine-tuning these models on normal industrial samples with SSL pretext tasks —such as masked patch reconstruction—promises to further sharpen their sensitivity to subtle deviations indicative of anomalies.

6.3 Explainability and Regulatory Requirements

Industrial VAD systems operate in regulated environments where explainability is a hard requirement. A detection decision must be accompanied by a localization map indicating which region triggered the anomaly call, and increasingly, regulatory frameworks require human-interpretable justifications. SSL-based reconstruction networks naturally produce pixel-level anomaly maps—regions with high reconstruction error correspond directly to anomalous regions—making them more interpretable than black-box classifiers (Liu et al., 2024). The integration of attention mechanisms and gradient-based explanation methods (GradCAM, Integrated Gradients) further enhances interpretability for SSL models that do not natively produce spatial outputs.

The broader integration of AI into industrial software systems also raises novel VAD challenges. Wang and colleagues (2025) demonstrated that large language models (LLMs) can automate end-to-end software testing pipelines—including test case generation, script authoring, and failure diagnosis—highlighting how AI-generated software artifacts introduce new categories of anomalies that traditional inspection methods cannot detect. This observation extends beyond code to AI-generated designs, process parameters, and documentation, all of which require anomaly detection capabilities as AI adoption deepens in manufacturing.

6.4 Real-Time Deployment Challenges

Deploying SSL-based VAD systems on real production lines imposes stringent latency constraints—often requiring sub-100ms inference per image at line speeds exceeding one item per second. Masked reconstruction models, while powerful, can be computationally expensive for high-resolution industrial images. Recent work on lightweight masked autoencoders, pruned vision transformers, and knowledge-distilled student models addresses this challenge, enabling SSL models to meet real-time requirements without sacrificing detection accuracy (Liu et al., 2024).

6.5 Open Challenges

Despite significant progress, several open challenges define the frontier of SSL for industrial VAD:

1. **Zero-shot generalization to novel defect types**: Current SSL methods excel at detecting deviations from normal but struggle to characterize the nature of novel defects. Addressing this requires integrating anomaly detection with anomaly characterization—potentially through vision-language models that can describe detected anomalies in natural language.
2. **Domain shift and model updating**: Industrial production conditions evolve—new product variants, changed materials, adjusted process parameters—which causes the distribution of normal data to drift. SSL models trained on historical normal data may become miscalibrated under domain shift. Continual self-supervised learning strategies that update the normal model without catastrophic forgetting are needed.
3. **Multi-modal anomaly detection**: Real industrial systems are inherently multi-modal—coordinating 2D vision, 3D geometry, acoustic emission, and process sensor data. Effective multi-modal SSL for VAD requires pretext tasks that exploit cross-modal consistency, a problem that remains largely unsolved.
4. **Industrial data privacy and federated learning**: Competition among manufacturers discourages sharing defect datasets, creating isolated data silos. Federated learning—where models are trained collaboratively without exchanging raw data—offers a path forward, but its combination with SSL for industrial VAD is still nascent (Khan et al., 2025).

---

7. Conclusion

Self-supervised learning has fundamentally reshaped the landscape of industrial visual anomaly detection, offering a principled path to learning robust defect detection models from the abundant normal data generated by production lines. By eliminating the dependence on large labeled anomaly datasets—a fundamental practical constraint in manufacturing—SSL enables zero-shot and few-shot anomaly detection that supervised approaches cannot match.

This review has examined the methodological landscape across five major SSL paradigms—contrastive learning, masked reconstruction, generative modeling, rotation prediction, and cross-modal pretext tasks—and traced their application across industrial domains from surface inspection to 3D quality control, semiconductor fabrication, and predictive maintenance. Key technological inflection points include the emergence of diffusion models for synthetic defect generation, the integration of SSL with digital twin platforms for predictive maintenance, and the transfer of vision foundation models to industrial inspection.

Looking ahead, the convergence of self-supervised representation learning, generative modeling, and digital twin technologies promises a new generation of industrial inspection systems: models that learn continuously from streaming production data, generate synthetic training examples on demand, and reason about anomalies across multiple data modalities—pushing the frontier of zero-defect manufacturing.

References

Deng, T., Li, Y., Liu, X., & Wang, L. (2023). Federated learning-based collaborative manufacturing for complex parts. *Journal of Intelligent Manufacturing*, 34(7), 3025–3038. https://doi.org/10.1007/s10845-022-01968-3

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. In *Advances in Neural Information Processing Systems* (NeurIPS) (pp. 2672–2680). Curran Associates, Inc.

He, K., Fan, H., Wu, Y., Xie, S., & Girshick, R. (2020). Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (CVPR) (pp. 9729–9738). IEEE. https://doi.org/10.1109/CVPR42600.2020.00980

Huang, H., Tang, J., Liu, T., & Huang, M.-L. (2026). Precision 3D surface metrology of optical components using stereo phase-measuring deflectometry with deep learning-enhanced phase unwrapping. *Proceedings of SPIE*, 0898. https://doi.org/10.1117/12.3093993

Huang, H., Yang, Y., & Zhu, Y. (2023). Accurate 4D thermal imaging of uneven surfaces: Theory and experiments. *International Journal of Heat and Mass Transfer*, 211, 124580. https://doi.org/10.1016/j.ijheatmasstransfer.2023.124580

Khan, T., Urfi Khan, T., Khan, A., Mollan, C., & Vilkonciene, I. M. (2025). Data-driven digital twin framework for predictive maintenance of smart manufacturing systems. *Machines*, 13(6), 481. https://doi.org/10.3390/machines13060481

Khan, Y., et al. (2025). A few-shot steel surface defect generation method based on diffusion models. *BMC Medical Informatics and Decision Making* (PMC). https://doi.org/10.1186/s12911-025-02912-9

Li, S., et al. (2024). Towards scalable 3D anomaly detection and localization: A benchmark via 3D anomaly synthesis and a self-supervised learning network (IMRNet). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (CVPR) (pp. 12456–12466). IEEE. https://doi.org/10.1109/CVPR52733.2024.01190

Li, Y., Lou, J., Cai, Z., Zheng, P., Wu, H., & Wang, X. (2024). An interactive gesture control system for collaborative manipulator based on Leap Motion Controller. *Advances in Mechanical Engineering*, 16(5), 16878132241253101. https://doi.org/10.1177/16878132241253101

Liu, J., Xie, G., Chen, R., Li, X., Wang, J., Liu, Y., Wang, C., & Zheng, F. (2024). A survey of deep Learning for industrial visual anomaly detection. *Artificial Intelligence Review*, 58, 178. https://doi.org/10.1007/s10462-025-11287-7

Liu, J., et al. (2024). Deep industrial image anomaly detection: A survey. *arXiv preprint arXiv:2401.01432*. https://doi.org/10.48550/arXiv.2401.01432

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning* (ICML) (pp. 8748–8763). PMLR.

Shen, Y., et al. (2025). AI-enhanced digital twins in maintenance: Systematic review, industrial challenges, and bridging research–practice gaps. *ScienceDirect*. https://doi.org/10.1016/j.promfg.2025.107634

Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems* (NeurIPS) (pp. 5998–6008). Curran Associates, Inc.

Wang, S., Yu, Y., Feldt, R., & Parthasarathy, D. (2025). Automating a complete software test process using LLMs: An automotive case study. In *2025 IEEE/ACM 47th International Conference on Software Engineering (ICSE)*. https://doi.org/10.1109/ICSE55347.2025.00211

Wang, X., et al. (2025). Generative and predictive AI for digital twin systems in manufacturing. *Frontiers in Artificial Intelligence*, 8, 1655470. https://doi.org/10.3389/frai.2025.1655470

Zhang, A., et al. (2025). Industrial multivariate time-series data anomaly detection incorporating attention mechanisms and adversarial training. *International Journal of Computer Integrated Manufacturing*, 38(12). https://doi.org/10.1080/0951192X.2025.2452985

Zhang, M., et al. (2025). AI-enabled defect detection in industrial products: A comprehensive survey, key insights and future research challenges. *ScienceDirect*. https://doi.org/10.1016/j.ijmachtools.2025.104960

Zhang, Y., et al. (2025). Latent diffusion models to enhance the performance of visual defect segmentation networks in steel surface inspection. *Sensors*, 24(18), 6016. https://doi.org/10.3390/s24186016

Zhu, Y., & Liu, Q. (2025). Toward transparent groundwater contamination risk forecasting: Integrating causal discovery and Bayesian graph neural networks. *Science of the Total Environment*, 998, 180233. https://doi.org/10.1016/j.scitotenv.2025.180233

Zhu, Y., & Liu, Q. (2026). Hybrid graph attention network-LSTM models for causal-aware supply chain forecasting. *Journal of Intelligent Manufacturing*. https://doi.org/10.1007/s10845-025-02782-3

Paper authored by Bao Tang