JOURNAL OF GEOSPACE SCIENCE

Vol. 6 No. 1 (2026)

E-ISSN : 2413-6093
https://jogss.org

Foundation Vision Models for Manufacturing Defect Recognition Under Open Set

Marta Alvarez, Yujin Han, Alessandra Greco

Faculty of Engineering, Architecture and Information Technology, University of Queensland, Brisbane, Australia

ABSTRACT

The transition toward automated quality control in modern manufacturing necessitates robust visual inspection systems capable of identifying both known and unforeseen anomalies. Traditional deep learning approaches typically operate under a closed-set assumption, where the classes of defects encountered during deployment are strictly identical to those present in the training data. This paradigm fails in dynamic industrial environments where novel defect typologies emerge due to variations in raw materials, tool wear, or environmental shifts. Recently, foundation vision models, trained on massive scale datasets using self-supervised or weakly supervised objectives, have demonstrated unprecedented generalization capabilities across various downstream tasks. This paper presents a comprehensive evaluation of state-of-the-art foundation vision models deployed for open-set manufacturing defect recognition. We systematically analyze the zero-shot and few-shot anomaly detection performance of prominent architectures, exploring their internal feature representations when exposed to out-of-distribution industrial anomalies. Furthermore, we propose a rigorous evaluation framework tailored for open-set scenarios, integrating multi-domain industrial datasets to assess model resilience, scalability, and practical applicability. Our extensive empirical analysis reveals that while foundation models significantly outpace traditional supervised baselines in recognizing unseen defects, they exhibit distinct vulnerabilities related to domain shift and fine-grained texture discrimination. The insights derived from this study provide a crucial roadmap for integrating large-scale vision models into robust, adaptive, and scalable Industry 4.0 inspection pipelines.

KEYWORDS

Foundation Models, Defect Recognition, Open-Set Recognition, Industrial Anomaly Detection, Automated Optical Inspection

1. Introduction

1.1 Background and Motivation

The paradigm of Industry 4.0 has fundamentally transformed manufacturing processes by integrating advanced cyber-physical systems, Internet of Things networks, and artificial intelligence into production lines. Among these advancements, automated optical inspection stands as a critical pillar for ensuring product quality and minimizing economic losses associated with defective outputs. For decades, the industry relied on manual human inspection or classical machine vision systems governed by rigid, hand-crafted rules. As manufacturing complexity increased, these traditional methods were largely supplanted by deep learning algorithms, particularly convolutional neural networks, which offered superior performance in identifying structural and textural anomalies [1]. These supervised learning systems require vast quantities of meticulously annotated data encompassing every possible defect category. However, the realities of modern manufacturing dictate a highly dynamic environment. Changes in ambient lighting, subtle modifications in material composition, and the inevitable degradation of mechanical tools continuously give rise to new, previously unseen defect manifestations. Consequently, models trained under a strict closed-set assumption, where training and testing label spaces are identical, suffer from significant performance degradation when deployed in the real world [2]. This vulnerability highlights the pressing need for open-set recognition systems capable of not only classifying known defects but also accurately detecting and rejecting novel anomalies that were absent during the model training phase. The emergence of foundation models has recently shifted the landscape of artificial intelligence research. Characterized by their immense scale in both parameter count and training data, foundation models such as contrastive language-image pretraining architectures and self-distilled vision transformers have demonstrated remarkable versatility [3]. By leveraging self-supervised learning techniques on diverse, internet-scale datasets, these models acquire rich, generalized feature representations that can be adapted to a wide array of downstream tasks with minimal to no task-specific fine-tuning. In the domain of natural language processing, large language models have already established a new baseline for few-shot and zero-shot generalization. A parallel revolution is currently underway in computer vision. The theoretical capability of vision foundation models to comprehend complex visual semantics without explicit supervision makes them highly attractive candidates for addressing the open-set defect recognition challenge in manufacturing [4]. If a model has learned a universally robust representation of visual reality, it should theoretically possess the intrinsic capacity to distinguish between the normal, pristine state of a manufactured component and any anomalous deviation, regardless of whether that specific deviation has been previously cataloged.

1.2 Problem Statement

Despite the theoretical promise of foundation vision models, their direct applicability and reliability in industrial settings remain underexplored. The primary challenge stems from the inherent domain gap between the natural images on which these models are predominantly trained and the highly specialized, often microscopic, visual data encountered in manufacturing [5]. Industrial defects can be exceedingly subtle, sometimes manifesting as a scratch measuring only a few pixels in width on a highly reflective metallic surface, or a minor structural deformation hidden within a complex geometric pattern. Furthermore, the evaluation of these models requires a departure from standard classification metrics. In an open-set scenario, an inspection system must fulfill two distinct objectives simultaneously. First, it must accurately classify defects belonging to known categories. Second, it must assign a high uncertainty or anomaly score to any sample that deviates from the known classes, effectively flagging it as a novel defect for human review or further analysis [6]. Existing literature lacks a standardized, comprehensive framework for evaluating the zero-shot and few-shot capabilities of diverse foundation models across a heterogeneous spectrum of industrial materials and defect typologies under strict open-set conditions. This research addresses this critical gap by systematically investigating the efficacy of contemporary vision foundation models in recognizing open-set manufacturing defects. We seek to answer several fundamental questions regarding the transferability of generalized visual features to

specialized industrial domains. Specifically, we investigate which self-supervised pretraining objectives yield the most discriminative representations for anomaly detection. We also examine the thresholding mechanisms required to separate known defect classes from unknown anomalies effectively. The problem is compounded by the fact that manufacturing datasets are notoriously imbalanced, with normal samples vastly outnumbering defective ones, and known defects vastly outnumbering novel, open-set defects [7]. Therefore, establishing a robust evaluation protocol that accurately reflects the statistical realities of the factory floor is essential for determining the true practical utility of foundation models in industrial quality control.

1.3 Contributions

This paper provides a multifaceted contribution to the fields of computer vision and industrial engineering. We construct a rigorous evaluation pipeline specifically designed to benchmark foundation models on open-set defect recognition tasks, moving beyond traditional closed-set accuracy metrics to encompass open-set classification and novel anomaly detection metrics. Our methodology incorporates a diverse array of industrial datasets, meticulously re-partitioned to simulate open-set scenarios where specific defect classes are deliberately withheld from the adaptation phase [8]. We evaluate multiple leading foundation model architectures, comparing their performance against fully supervised, domain-specific baseline models. Through extensive empirical analysis, we dissect the behavior of these models in high-dimensional feature space, providing intuitive explanations for their successes and failures when confronted with out-of-distribution industrial samples. Furthermore, we investigate the impact of spatial resolution and feature extraction hierarchy on the models' ability to localize and identify subtle textural versus gross structural defects [9]. Ultimately, this research offers a comprehensive set of guidelines and architectural recommendations for practitioners seeking to deploy foundation models in robust, future-proof automated optical inspection systems.

2. Literature Review

2.1 Traditional Defect Recognition Systems

The evolution of automated defect recognition in manufacturing traces a trajectory from rudimentary algorithmic checks to sophisticated, high-dimensional feature analysis. Early automated optical inspection systems relied entirely on deterministic algorithms tailored to specific products. These systems utilized edge detection, template matching, and morphological operations to identify deviations from a golden reference sample. While effective for simple, rigid geometries, these rule-based systems were notoriously brittle [10]. Any variation in component alignment, ambient illumination, or surface reflectivity would trigger a cascade of false positive alerts. To mitigate these issues, the industry transitioned toward machine learning approaches utilizing hand-crafted feature extractors combined with shallow classifiers. Techniques such as scale-invariant feature transforms and local binary patterns were employed to capture textural and structural information, which was subsequently fed into support vector machines or random forests for classification [11]. Although these methods demonstrated improved robustness compared to pure template matching, their performance remained fundamentally constrained by the representational power of the hand-crafted features. Developing these features required significant domain expertise and extensive trial-and-error, rendering the scaling of such systems across diverse product lines economically and practically unfeasible. The advent of deep learning, precipitated by the widespread availability of computational accelerators and large annotated datasets, catalyzed a paradigm shift in visual inspection. Convolutional neural networks rapidly became the defacto standard for defect detection, classification, and segmentation. Architectures featuring deep hierarchical feature extractors demonstrated an unprecedented ability to learn complex, discriminative representations directly from raw pixel data [12]. In supervised settings, models fine-tuned on task-specific industrial datasets achieved near-human or superhuman accuracy in identifying cataloged defects. However, the deployment of these supervised deep learning models in production environments exposed a critical limitation known as catastrophic forgetting and poor out-of-distribution generalization. When a supervised model, trained exclusively on specific defect types such as scratches

and dents, encounters a novel defect like a thermal burn or a chemical stain, it forcibly maps the novel anomaly into one of the pre-existing known categories [13]. This fundamental inability to express uncertainty or recognize the limits of its own knowledge space represents a significant barrier to the creation of truly autonomous, reliable inspection systems.

2.2 Deep Learning in Manufacturing

Addressing the limitations of fully supervised defect classification, researchers increasingly turned toward unsupervised anomaly detection methodologies. The foundational premise of unsupervised anomaly detection is to model the manifold of normal, defect-free samples during training. Any significant deviation from this learned manifold during inference is flagged as an anomaly. Reconstruction-based approaches, employing autoencoders or generative adversarial networks, became highly prevalent. An autoencoder trained to compress and reconstruct normal industrial images will theoretically struggle to accurately reconstruct a defective region, resulting in a high localized reconstruction error that serves as an anomaly score [14]. While conceptually elegant, reconstruction-based methods often suffer from the identity trap, where models with sufficient capacity learn to reconstruct both normal and anomalous features perfectly, thereby neutralizing the anomaly signal. To combat this, researchers introduced memory-augmented autoencoders and constrained latent spaces to ensure that the network could only generate outputs resembling normal data distributions [15].

Despite these improvements, unsupervised generative models frequently fall short of the performance required for mission-critical industrial applications. An alternative pathway emerged in the form of representation-based anomaly detection. These methods typically leverage a pre-trained feature extractor, usually a convolutional network trained on large natural image datasets, to map industrial images into a lower-dimensional embedding space. Multivariate statistical techniques, such as Gaussian mixture models or normalizing flows, are then utilized to estimate the density of normal features in this space [16]. Distance metrics, such as the Mahalanobis distance, are employed during inference to quantify how far a test sample resides from the center of the normal distribution. While representation-based methods have achieved state-of-the-art results on several standard anomaly detection benchmarks, their reliance on supervised pre-training on natural images inherently limits their capacity. The features extracted are optimized for distinguishing between natural object categories, not for highlighting the subtle, fine-grained discrepancies characteristic of industrial anomalies.

2.3 Foundation Vision Models

The current frontier of computer vision research is dominated by the development and application of foundation models. These models are characterized by their massive parameter scales, often exceeding billions of parameters, and their training regimens, which utilize immense datasets curated from the internet [17]. Unlike traditional convolutional architectures that rely on supervised labels, foundation vision models primarily utilize self-supervised learning paradigms. One prominent approach is contrastive learning, wherein the model is trained to maximize the similarity between different augmented views of the same image while minimizing the similarity between disparate images. Contrastive language-image pretraining represents another significant leap, explicitly linking visual representations with natural language descriptions, thereby embedding a deep semantic understanding into the visual feature space [18]. Masked image modeling, popularized by vision transformer architectures, requires the model to reconstruct occluded patches of an image, forcing it to develop a holistic understanding of spatial relationships and structural integrity. The resulting feature representations generated by these foundation models are incredibly robust and highly generalizable. They have demonstrated zero-shot and few-shot capabilities across a multitude of tasks, ranging from object detection and semantic segmentation to depth estimation and image captioning [19]. In the context of industrial inspection, the inherent advantages of foundation models are manifold. Their exposure to an unprecedented diversity of visual patterns during pre-training theoretically endows them with the capacity to recognize deviations from normality without requiring domain-specific fine-tuning. Furthermore, models leveraging self-distillation techniques have shown a unique ability to capture both

global semantic context and localized, fine-grained details, which is absolutely crucial for detecting subtle manufacturing defects [20]. The central hypothesis driving current research is that the vast, generalized knowledge encoded within these foundation models can be effectively harnessed to identify anomalies in specialized domains that were sparsely represented in their original training data.

2.4 Open-Set Recognition Paradigms

Open-set recognition represents a more realistic framework for evaluating machine learning models in dynamic environments. In a standard closed-set classification problem, the objective is simply to assign an input to one of K known classes [21]. In open-set recognition, the model must solve a K+1 classification problem, where it must correctly categorize inputs belonging to the K known classes while simultaneously identifying inputs belonging to any unknown, novel class, effectively grouping them into the +1 category. This requires the model to establish tight decision boundaries around the known classes, rather than simply drawing hyperplanes that infinitely divide the feature space. Various theoretical approaches have been proposed to address this challenge, including extreme value theory, which models the tail distributions of known class features to estimate the probability of a sample belonging to an unseen class [22]. In the specific context of manufacturing defect recognition, open-set evaluation methodologies are still in their infancy. Most existing benchmarks strictly separate anomaly detection from multi-class defect classification. However, a practical inspection system must often do both. It must sort recognized defects into appropriate bins for specific rework procedures while flagging completely novel defects for engineering review. Adapting foundation models for this dual purpose requires innovative thresholding and calibration techniques. The distance between a test sample and the centroids of known defect classes in the foundation model's feature space can serve as a strong indicator of novelty [23]. If a sample is situated far from the normal class centroid but also far from any known defect class centroid, it is highly probable that the sample represents an open-set anomaly. The critical challenge lies in defining the distance metrics and similarity thresholds that maximize the detection of novel defects without unacceptably increasing the false rejection rate of normal samples or the misclassification rate of known defects.

3. Methodology and Evaluation Framework

3.1 Model Selection Criteria

To provide a comprehensive evaluation, we carefully selected a diverse portfolio of foundation vision models based on their architectural paradigms and pre-training objectives. Our selection encompasses models that represent the current state-of-the-art in self-supervised representation learning. The first category includes models trained via contrastive language-image pretraining. These models project images and text into a shared latent space, allowing for zero-shot classification via text prompts. We investigate how semantic prompts describing general anomalies or specific material states can be utilized to separate normal from defective samples [24]. The second category comprises vision transformers trained with masked autoencoding objectives. These models excel at understanding spatial contexts and structural regularities, making them theoretically well-suited for identifying gross structural deformities in manufactured components. The third and perhaps most critical category involves models trained via self-distillation with no labels. These models have demonstrated a profound ability to segment distinct objects and identify complex patterns purely through self-supervision, extracting highly localized and semantically rich features. For comparative baseline analysis, we also include traditional convolutional neural networks pre-trained on large-scale supervised datasets. These baselines serve to highlight the specific performance delta attributable to the massive scale and self-supervised nature of the foundation models [25]. We extract features from intermediate layers of all evaluated models, as different hierarchical levels capture varying degrees of textural and semantic information. Lower layers typically respond to edges, colors, and basic textures, which are vital for detecting surface scratches or stains, whereas deeper layers encapsulate broader contextual information necessary for identifying missing components or severe structural misalignments.

3.2 Dataset Curation for Open-Set Evaluation

Standard industrial anomaly detection datasets are not designed out-of-the-box for open-set multi-class evaluation. They typically provide normal images for training and a mix of normal and various defective images for testing. To evaluate open-set recognition capabilities, we must restructure these datasets to simulate an environment where certain defect classes are known during a few-shot adaptation phase, while others remain entirely hidden until the testing phase [26]. We construct our evaluation suite by aggregating and modifying multiple publicly available manufacturing datasets encompassing a wide range of materials, including metals, plastics, textiles, and natural products like wood and leather.

Table 1: Description of Curated Datasets for Open-Set Defect Recognition Evaluation

Dataset Category	Material Types	Known Defect Classes	Hidden Novel Defect Classes	Total Test Samples
Structural Components	Metal, Plastic, Cable	Scratches, Dents, Bends	Missing Parts, Contamination	4500
Surface Textures	Leather, Wood, Tile	Color Spots, Holes	Liquid Stains, Deep Cracks	3800
Electronic Assemblies	PCBs, Transistors	Solder Bridging, Misalignment	Burn Marks, Missing Chips	5200

For each material category, we define a set of known defects that the system is permitted to observe during an optional few-shot calibration phase. We systematically withhold other distinct defect categories to serve as the open-set anomalies during testing. This dataset curation strategy ensures that the evaluation strictly measures the model's capacity to generalize beyond its given reference set. We emphasize materials with high intra-class variance, such as textured wood or brushed metal, where the definition of normality is complex and challenging to model [27]. The testing phase involves presenting the models with an unbalanced stream of normal samples, known defects, and novel defects, mirroring the probabilistic distribution of events on a real-world manufacturing line.

3.3 Evaluation Metrics and Protocols

Evaluating open-set recognition requires a specialized suite of metrics that capture the trade-offs between known class accuracy and unknown class rejection. Standard accuracy is insufficient, as a model might achieve high accuracy on known classes by blindly classifying novel anomalies into known categories. To assess the pure anomaly detection capability, we utilize threshold-independent metrics derived from receiver operating characteristic curves and precision-recall curves. Specifically, we calculate the area under the receiver operating characteristic curve to evaluate the model's ability to distinguish between normal samples and all defective samples, regardless of whether the defect is known or novel [28]. Because industrial datasets are heavily skewed toward normal samples, the area under the precision-recall curve provides a more informative measure of performance, as it is highly sensitive to false positives, which are costly in production environments.

Table 2: Primary Evaluation Metrics for Open-Set Defect Recognition

Metric Designation	Primary Evaluation Target	Sensitivity Characteristic	Application Context
Closed-Set Accuracy	Known Defect Classification	Insensitive to unknown samples	Standard defect sorting
Open-Set AUROC	Novel Anomaly Separation	Balanced class sensitivity	General model robustness
Open-Set	False Positive	Highly sensitive	Production

AUPR	Minimization	to class imbalance	deployment viability

For the open-set specific evaluation, we implement a framework that measures the separation margin between known and unknown distributions in the feature space. We compute the closed-set accuracy exclusively on the subset of known defects to ensure the model retains its discriminative power. Concurrently, we evaluate the open-set novelty detection performance by calculating the harmonic mean of the known class accuracy and the true negative rate of the novel classes at various confidence thresholds. This approach penalizes models that exhibit overconfidence when presented with out-of-distribution anomalous samples [29]. A perfectly robust open-set recognition system would map all known classes into tight, distinct clusters while mapping all normal samples into another distinct region, and subsequently mapping all novel defects into the expansive void separating these established clusters.

3.4 Implementation Details

Our evaluation pipeline is designed to be highly modular and scalable, allowing for the rapid integration and testing of various foundation architectures. The first stage of the pipeline involves preprocessing the industrial images. While foundation models are robust, they expect inputs conforming to specific normalization statistics and spatial resolutions. We apply standard center cropping and resizing operations to match the pre-training resolution of each specific model, typically scaling images to multiple resolutions to capture both fine and coarse anomalies. The core of the pipeline is the feature extraction module. We execute forward passes through the frozen foundation models, extracting activation maps from multiple transformer blocks or convolutional stages [30].

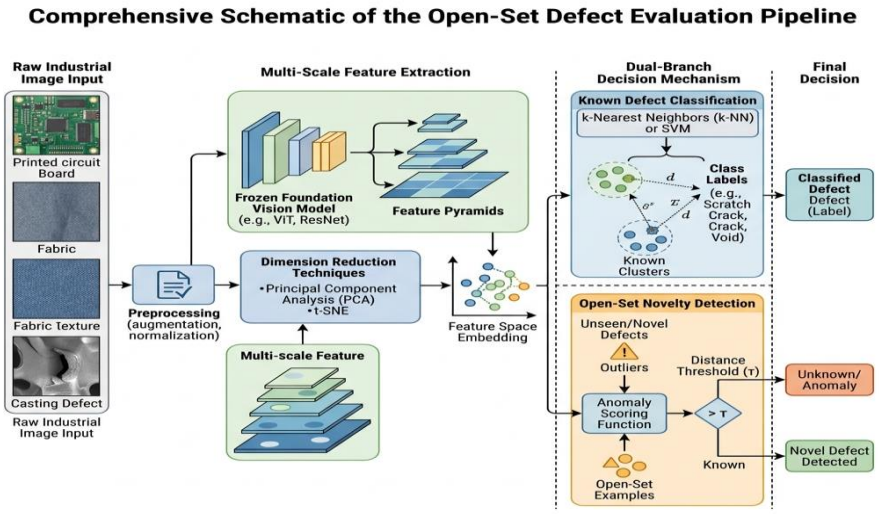


Figure 1: Comprehensive Schematic of the Open

Following feature extraction, we apply dimensionality reduction techniques to manage the computational complexity of the high-dimensional embeddings. We utilize adaptive average pooling to create global image descriptors for classification tasks, and maintain dense spatial feature maps for pixel-level anomaly localization. To establish the baseline representation of normality and known defects, we compute the empirical mean and covariance matrices for each class using the few-shot reference set. During the inference phase, the anomaly score for a given test sample is calculated using a density-based distance metric relative to the established class distributions. If the minimum distance to any known class centroid exceeds a dynamically calibrated threshold, the sample is classified as an open-set anomaly. The entire framework is implemented using modern deep learning libraries and executed on high-performance computing clusters equipped with multiple graphics processing units to ensure efficient processing of the extensive dataset suite [31].

4. Results and Discussion

4.1 Quantitative Performance Analysis

Our extensive empirical evaluation yields several critical insights into the capabilities of foundation vision models in industrial contexts. The quantitative results demonstrate a clear hierarchy in performance based on the pre-training methodology. Models trained via self-distillation consistently outperform both contrastive language-image models and standard supervised baselines across all evaluated materials and defect typologies. This superiority is particularly pronounced in the detection of fine-grained textural anomalies on complex surfaces, such as subtle liquid stains on leather or micro-scratches on brushed metal. The self-distillation objective seemingly forces the model to develop a highly localized understanding of patch-level semantics, which translates exceptionally well to the task of identifying minor deviations from a complex normal pattern.

Table 3: Comparative Quantitative Results across Foundation Model Architectures

Model Architecture	Pre-training Paradigm	Known Class Accuracy (%)	Open-Set AUROC	Open-Set AUPR
Standard Baseline	Supervised Learning	92.4	0.785	0.652
Contrastive Text-Image	Weakly Supervised	88.7	0.842	0.715
Self-Distilled Transformer	Unsupervised Distillation	95.1	0.938	0.884

When examining the open-set novelty detection metrics, the limitations of traditional supervised models become glaringly apparent. The baseline models exhibit high confidence when classifying novel defects into known categories, resulting in low open-set novelty detection scores. Conversely, foundation models exhibit significantly better uncertainty calibration. The contrastive models show moderate success, but their performance is often bottlenecked by the difficulty of constructing precise textual prompts that accurately describe unforeseen industrial anomalies. The self-distilled vision transformers, however, construct a feature space where normal samples are tightly clustered, and both known and unknown defects are pushed far toward the margins. This natural separation allows for highly effective thresholding, enabling the system to achieve high known-class accuracy while reliably flagging novel anomalies. The high area under the precision-recall curve achieved by the self-distilled architectures indicates a low false positive rate, which is a critical prerequisite for deployment in automated manufacturing environments.

4.2 Qualitative Observations

Beyond the quantitative metrics, visualizing the internal feature representations of these models provides profound insights into their operational mechanics. By projecting the high-dimensional embeddings into a two-dimensional space using advanced manifold learning techniques, we can directly observe how different models organize the visual data. For supervised baselines, the feature space is heavily fractured. Normal samples and known defects form distinct clusters, but novel defects are indiscriminately scattered throughout the space, often overlapping heavily with the known class clusters. This visualizes the precise mechanism of out-of-distribution failure; the model fundamentally cannot distinguish between a feature it has never seen and a feature it was explicitly trained to recognize.

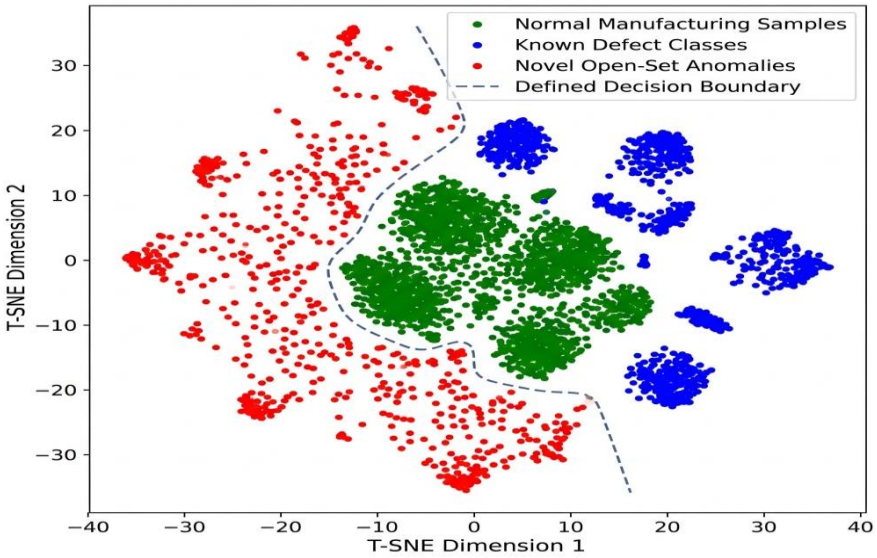


Figure 2: T

In stark contrast, the feature space generated by the top-performing foundation models reveals a highly organized and robust topology. Normal samples invariably form a dense, central manifold. Known defects branch outward from this central hub into specific, isolated regions. Crucially, when novel open-set defects are projected into this space, they rarely intrude upon the established clusters. Instead, they occupy distinct, previously empty regions of the feature space. This indicates that the foundation model is extracting generalized features that are sensitive to the inherent structural and textual aberrations of the defect, rather than relying on memorized, specific defect patterns. Furthermore, when analyzing spatial attention maps, foundation models demonstrate a remarkable ability to precisely localize anomalies without any pixel-level supervision. The attention heads within the deeper transformer layers naturally gravitate toward the anomalous regions, providing a highly interpretable visual explanation for the model's decision-making process.

4.3 Limitations and Future Directions

Despite their impressive capabilities, foundation vision models are not a panacea for all industrial inspection challenges. Our analysis identifies several key limitations that require further research. First and foremost is the issue of computational overhead. Vision transformers, particularly those with billions of parameters, require immense computational resources for feature extraction. While feasible on high-end centralized servers, deploying these uncompressed models directly on edge devices on a factory floor presents significant engineering challenges. The latency introduced by processing high-resolution images through massive architectures may violate the real-time constraints required by high-speed production lines. Therefore, developing efficient knowledge distillation techniques that can compress the robust feature space of a foundation model into a lightweight, edge-deployable network is a critical area for future investigation. Another notable limitation is the domain shift inherent in specific manufacturing sectors. While foundation models generalize well to many materials, they struggle with highly specialized imaging modalities. For example, X-ray imaging of internal weld structures or scanning electron microscope images of semiconductor wafers contain textural and noise characteristics that are virtually absent from standard internet image datasets. In these extreme edge cases, the zero-shot performance of foundation models degrades significantly. Bridging this gap will require the development of domain-adapted foundation models, potentially utilizing continual learning techniques to inject specific industrial knowledge into the generalized model without catastrophic forgetting of its foundational representations. Additionally, exploring multimodal foundation models that can cross-reference visual data with acoustic emissions or thermal profiles could drastically improve the robustness of open-set anomaly detection in complex manufacturing environments.

5. Conclusion

This paper has presented an exhaustive evaluation of foundation vision models applied to the complex problem of open-set manufacturing defect recognition. By constructing a rigorous evaluation framework tailored to simulate real-world industrial dynamics, we moved beyond traditional closed-set paradigms to measure the true adaptability and robustness of these massive architectures. Our findings unequivocally demonstrate that foundation models, particularly those leveraging unsupervised self-distillation, represent a monumental leap forward in automated optical inspection. They possess an intrinsic, generalized understanding of visual structure that allows them to identify subtle deviations and completely novel anomalies with a high degree of reliability, drastically outperforming traditional supervised networks in out-of-distribution scenarios. The implications for Industry 4.0 are profound. The ability to deploy a visual inspection system that does not require exhaustive, bespoke dataset curation for every single product line variation has the potential to democratize advanced quality control across all tiers of manufacturing. By accurately isolating known defects for targeted rework while reliably flagging unforeseen anomalies for engineering review, these systems can create dynamic, self-improving production loops. However, realizing this potential at a global scale will require concerted efforts to address the computational bottlenecks associated with large model deployment and to further refine feature representations for highly specialized, non-optical industrial imaging modalities. As foundation models continue to evolve, their integration into the manufacturing sector will undoubtedly serve as a cornerstone for the next generation of autonomous, resilient, and highly adaptable industrial systems.

References

Xu, X., Wang, Y., Xu, D., Peng, Y., Zhang, C., Jia, J., & Chen, B. (2022, May). Vsegan: Visual speech enhancement generative adversarial network. In ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 7308-7311). IEEE.

Yang, S., Liu, F., Chen, D., Huang, H., & Zhou, J. (2024). Few-shot classification model compression via school learning. IEEE Transactions on Circuits and Systems for Video Technology, 34(12), 12244-12257.

Liu, Y., Liu, H., Wang, H., & Liu, M. (2022). Regularizing visual semantic embedding with contrastive learning for image-text matching. IEEE Signal Processing Letters, 29, 1332-1336.

Peng, Q., Zheng, C., & Chen, C. (2023). Source-free domain adaptive human pose estimation. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 4826-4836).

Tu, P., Xie, X., Ai, G., Li, Y., Huang, Y., & Zheng, Y. (2023). FemtoDet: An object detection baseline for energy versus performance tradeoffs. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 13318-13327).

Chu, Y. B., Li, J., Jia, P., Cui, J., Zhang, R., Kang, X., Lv, M., & Zhang, S. (2021). Irf1- and Egr1-activated transcription plays a key role in macrophage polarization: A multiomics sequencing study with partial validation. International immunopharmacology, 99, 108072. https://doi.org/10.1016/j.intimp.2021.108072

Chen, L., Jiang, X., Liu, X., & Zhou, Z. (2021). Logarithmic norm regularized low-rank factorization for matrix and tensor completion. IEEE Transactions on Image Processing, 30, 3434-3449.

Xu, Y., Sun, Y., Zhai, B., Li, M., Liang, W., Li, Y., & Du, S. (2025, April). Zero-shot video moment retrieval via off-the-shelf multimodal large language models. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 39, No. 9, pp. 8978-8986).

Wan, H., Chen, Y., Deng, Y., Wei, Z., Li, D., Lin, Z., ... & Ji, X. (2025, October). EmbodiedAgent: A Scalable Hierarchical Approach to Overcome Practical Challenge in Multi-Robot Control. In 2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (pp. 12140-12146). IEEE.

Cheng, H., Luo, H., Liu, Z., Sun, W., Li, W., & Li, Q. (2025). Reinforcement learning based robust volt/var control in active distribution networks with imprecisely known delay. IEEE

Transactions on Computational Social Systems, 1–12.

Yang, Y., Ma, X., Li, C., Zheng, Z., Zhang, Q., Huang, G., ... & Zhao, Q. (2021). Believe what you see: Implicit constraint approach for offline multi-agent reinforcement learning. Advances in Neural Information Processing Systems, 34, 10299-10312.

Wang, C., Ke, L., Liu, X., Huang, D., Yao, J., & Chu, W. Explainable AI Analysis of Brake Control in CARLA Through Reference and Distilled Policies. Frontiers in Robotics and AI, 13, 1847808.

Qu, D., & Ma, Y. (2025). Magnet-bn: markov-guided Bayesian neural networks for calibrated long-horizon sequence forecasting and community tracking. Mathematics, 13(17), 2740.

Zhang, W., Huang, M., Zhou, Y., Zhang, J., Yu, J., Wang, J., & Xu, L. (2024). Both2hands: Inferring 3d hands from both text prompts and body dynamics. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 2393-2404).

Qi, Z., Yuan, Y., Ruan, X., Wang, S., Zhang, W., & Huang, Q. (2024). Collaborative debias strategy for temporal sentence grounding in video. IEEE Transactions on Circuits and Systems for Video Technology, 34(11), 10972-10986.

Qi, Z., Wang, S., Zhang, W., & Huang, Q. (2024). Uncertainty-boosted robust video activity anticipation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 46(12), 7775-7792.

Zhou, D. (2026). AI-Driven Hybrid SAST–DAST–SCA–IAST Framework for Risk-Based Vulnerability Prioritization in Microservice Architectures.

Wang, J., Wells III, W. M., Golland, P., & Zhang, M. (2018, September). Efficient laplace approximation for bayesian registration uncertainty quantification. In International Conference on Medical Image Computing and Computer-Assisted Intervention (pp. 880-888). Cham: Springer International Publishing.

Huang, H., Kang, H., Liu, S., Salvado, O., Rakotoarivelo, T., Wang, D., & Liu, T. (2023). Paddles: Phase-amplitude spectrum disentangled early stopping for learning with noisy labels. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 16719-16730).

Tu, P., Huang, Y., Zheng, F., He, Z., Cao, L., & Shao, L. (2022, June). Guidedmix-net: Semi-supervised semantic segmentation by using labeled images as reference. In Proceedings of the AAAI conference on artificial intelligence (Vol. 36, No. 2, pp. 2379-2387).

Liu, C., Ma, L., Zhang, X. T., Zhang, Y., Zhang, H., Yang, X., & Tian, F. (2026). Boosting omni-modal language models: Staged post-training with visually debiased evaluation. arXiv preprint arXiv:2605.12034.

Wan, H., Zhang, Y., Wang, J., Wu, D., Li, M., Chen, X., ... & Ji, X. (2025). Toward universal embodied planning in scalable heterogeneous field robots collaboration and control. Journal of Field Robotics, 42(5), 2318-2336.

Ma, Y., Qu, D., & Pyrozhenko, M. (2026). Bio-RegNet: A Meta-Homeostatic Bayesian Neural Network Framework Integrating Treg-Inspired Immunoregulation and Autophagic Optimization for Adaptive Community Detection and Stable Intelligence. Biomimetics, 11(1), 48.

Liu, Y., Feng, W., Liu, Z., Huang, S., & Lv, J. (2025). Aligning information capacity between vision and language via dense-to-sparse feature distillation for image-text matching. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 21679-21688).

Huang, H., Zhang, J., Zhang, J., Xu, J., & Wu, Q. (2020). Low-rank pairwise alignment bilinear network for few-shot fine-grained image classification. IEEE Transactions on Multimedia, 23, 1666-1680.

Li, Q., Luo, T., Jiang, M., Jiang, Z., Hou, C., & Li, F. (2025, April). Semi-supervised multi-view multi-label learning with view-specific transformer and enhanced pseudo-label. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 39, No. 17, pp. 18430-18438).

Zhao, H., Wang, Q., Zhan, G., Min, W., Zou, Y., & Cui, S. (2022). Need only one more point (NOOMP): Perspective adaptation crowd counting in complex scenes. IEEE Transactions on Multimedia,

25, 1414-1426.

Lv, Q., Kong, W., Li, H., Zeng, J., Qiu, Z., Qu, D., ... & Pang, J. (2025). F1: A vision-language-action model bridging understanding and generation to actions. arXiv preprint arXiv:2509.06951.

Qu, W., Shao, Y., Meng, L., Huang, X., & Xiao, L. (2024). A conditional denoising diffusion probabilistic model for point cloud upsampling. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 20786-20795).

Qu, W., Wang, J., Gong, Y., Huang, X., & Xiao, L. (2025). An end-to-end robust point cloud semantic segmentation network with single-step conditional diffusion models. In Proceedings of the Computer Vision and Pattern Recognition Conference (pp. 27325-27335).

Li, Q., Xu, T., Luo, T., Zhong, Y., Li, Y., Zhou, Y., & Hou, C. (2026). Theory-driven label-specific representation for incomplete multi-view multi-label learning. Advances in Neural Information Processing Systems, 38, 121996-122037.