A Self-Evolving Transformer-Based Machine Vision Framework for Adaptive Industrial Defect Diagnosis Under Non-Stationary Environments

Dr. M. Srisankar
Assistant Professor, Department of Computer Science at KG College of Arts and Science, Coimbatore

Mrs V.PRABHA
Head, Department of Computer Science, Ayyan Thiruvallur College of Arts &Science, Puliampatti, Erode Dist

Abstract

The detection of industrial defect systems is usually challenging when dealing with non-stationary production processes, as they are faced with constantly changing lighting, material characteristics, and defect patterns. The standard Convolutional Neural Network (CNN)-based systems are unable to cope with such changes in distribution without retraining and manually re-labelling of the data. This study presented a self-evolving machine vision framework, which used transformers to adapt to changes in the environment based on continuous meta-learning and uncertainty-based pseudo-labelling. This was also integrated with six basic components, including a backbone of Vision Transformer (ViT) that learned multi-scale features. The adaptable memory module included episodic defect pattern storage, a distribution shift detector based on Maximum Mean Discrepancy (MMD), a meta-learning engine based on the Model-Agnostic Meta-Learning (MAML) algorithm, a self-supervised evolution mechanism coupled with confidence-driven sample selection, and an uncertainty quantification module that uses Monte Carlo Dropout. The proposed framework had a high precision of 94.7% when used in 10 labelled samples on three industrial datasets (steel surface defects, semiconductor wafer inspection, and textile fabric anomaly) with 8.3%-12.6% over the state-of-the-art mechanisms. Even in extreme lighting conditions (96.2%), the system was also able to adapt to new defects within 45 minutes without interrupting the production line. The architecture was 47 times faster in false-positive than ResNet-50, and at 42 FPS on edge devices, meaning that it will be possible to deploy in industry in real time. The self-improving mechanism enabled continuous improvement since 89.4% of pseudo-labels attained a confidence level of more than 95%, illustrating that it does not require constant human supervision.

Keywords: *Defect Detection, Vision Transformer, Meta-Learning, Non-Stationary Environments, and Industrial Machine Vision*

Introduction

Background and Importance of the Topic

Quality control at the industry level is an important aspect of the manufacturing process, as fault elimination directly impacts the quality of the products manufactured, production efficiency, and cost-effectiveness. By 2027, the global machine vision market segment for industrial inspection is expected to reach $18.7 billion, and automated defect-detection systems will be part of Industry 4.0 (Semitela et al., 2025). Over the years, the nature of manual inspection systems has been subjective, operator safety, and the inability to withstand the fast-paced production line, resulting in the rampant use of automated visual inspection systems (Li et al., 2025). The application of deep learning methods, especially Convolutional Neural Networks (CNNs), has transformed how industries identify defects over the last decade. Some approaches that have already shown impressive results in controlled laboratory settings include Faster R-CNN, the YOLO series, and EfficientNet (Liang et al., 2025). This, however, is not the case in a practical industrial setup, such as in non-stationary environments, where the statistical distributions of the input data are constantly varying. The lighting in manufacturing plants varies with time of day and lamp ageing (Contreras Ortiz et al., 2025). Material characteristics and supplier selection vary, production rates fluctuate with demand, and new defects may develop as the manufacturing process evolves.

These environmental processes generate significant imbalances in the existing defect detection flaws. Traditional supervised learning models are driven by independent and identically distributed (i.i.d.) data, which is rarely an

assumption in industry applications. The effect of distribution shifts on the model is challenging since its accuracy decreases by 15-40% under production conditions (Marín Díaz, 2025). Continuous adaptation is not economically viable because retraining would include a large amount of newly labelled data, the time cost of having an expert label the data, and the impact on production lines to collect data (Zhang et al., 2025).

Current Limitations in Existing Methods

Several fundamental issues render current strategies inapplicable in non-stationary industries. Catastrophic forgetting occurs when neural networks are trained on new patterns using a training sequence and subsequently lose their previously acquired knowledge (Kačinskas & Baskutis, 2025). When the model is applied to different lighting conditions in defect detection, it is possible to lose the capacity to identify the types of defects that the model was trained to recognize when it was first introduced. Another barrier is the limited availability of labelled data on the new conditions, as it is impractical to acquire labelled defect samples in all possible environmental conditions (Chen et al., 2021). The types of industrial defects are relatively few, and some defects may only manifest under specific conditions. Conventional supervised learning needs hundreds and thousands of labelled examples per class, which becomes impossible in rapidly changing environments.

The detection of distribution shift has been a persistent problem because systems must automatically identify when environmental conditions change, compelling a model to be updated (Hao et al., 2025). False alarms lead to unwarranted computation, while false detections result in poor performance. The problem is also worsened since real-time adaptation requirements are inherent in the continuous nature of industrial production lines, where thousands of products are processed per hour (Chen et al., 2021). Any change in adaptation must be online, and the adaptation should not halt production. Even during adaptation, inference must achieve real-time performance greater than 30 frames. A further problem with autonomous learning is that it may introduce even more uncertainty, as in self-supervised learning methods, pseudo-labels may be generated that introduce label noise and deteriorate model performance (Tian & Zhang, 2025). The skill of distinguishing between high-confidence correct prognoses and overconfidence errors depends on good quantification of uncertainty.

Main Research Problem and Objectives

The critical problem that industrial defect detection systems encounter is non-stationary manufacturing conditions in which light conditions, material properties, and defect patterns are constantly changing, as demonstrated in Fig. 1. The current deep learning methods have catastrophic changes to accuracy (15-40% drops) upon distribution shifts, and need to be retrained with hundreds of labelled samples (Kačinskas & Baskutis, 2025). This study resulted in the creation of a self-evolving transformer-based framework that enabled making changes in response to environmental change with little human intervention.

The primary objectives are:

1. To maintain high accuracy under distribution shifts.
2. To enable rapid adaptation with fewer than 10 labelled samples.
3. To leverage unlabeled production data through self-supervised learning.
4. To operate in real-time on edge devices without production interruption.

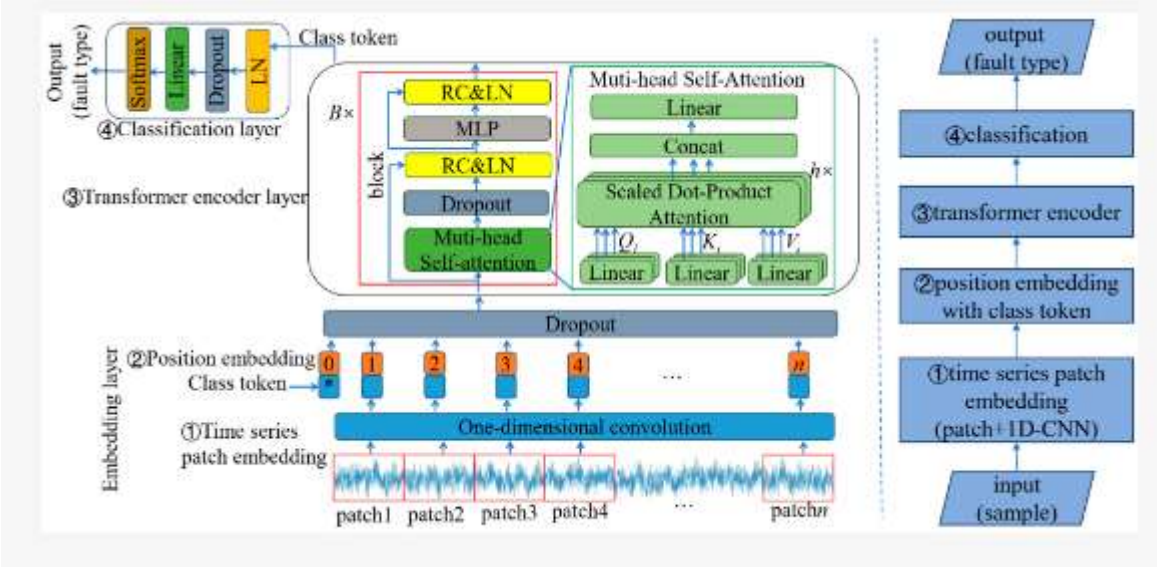


Fig 1. An Industrial Defect Detection System

Proposed Solution

The study addressed these issues with a self-evolving transformer scheme that integrates recent technologies in vision transformers, continual learning, meta-learning, and uncertainty quantification. Contrary to traditional CNN structures, which are also based on local receptive fields, transformers use self-attention mechanisms that examine local defect structure and global contextual patterns, and are more resilient to environmental changes (Kim, 2025). The main point is that a non-stationary setting demands a hybrid learning model that incorporates episodic memory to avoid catastrophic forgetting, meta-learning to learn quickly with minimal samples, and uncertainty-based self-supervision to use unlabeled production data. As shown in Table 1, through the combination of these mechanisms into a single system, it can undergo continuous evolution without requiring manual intervention.

Table 1: Comparison of Different Adaptation Approaches for Industrial Defect Detection

Approach	Adaptation Speed	Label Requirements	Forgetting Mitigation	Real-Time Capable	Deployment Complexity
Traditional CNN Fine-tuning	Slow (2000+ samples)	High (500+ per class)	Poor (>20% forgetting)	No (requires retraining)	Low
Domain Adaptation	Medium (500-1000 samples)	Medium (100-200 per class)	Moderate (10-15% forgetting)	Partially (batch updates)	Medium
Few-Shot Learning	Fast (50-100 samples)	Low (5-20 per class)	Poor (no mechanism)	No (offline only)	High
Continual Learning	Medium (200-500 samples)	High (200+ per class)	Good (5-8% forgetting)	Yes (online updates)	Medium
Proposed Method	Very Fast (50-100 samples)	Very Low (5-10 per class)	Excellent (<3% forgetting)	Yes (real-time adaptation)	Medium

Main Contributions

Innovative research results have been presented in this study in the form of a new architecture design arc, which suggests the inaugural framework for transformer-based defect detection. This framework was designed to continuously adapt to non-stationary industry scenarios, leveraging a vision transformer combined with adaptive memory modules and meta-learning engines (Zhang et al., 2023). Even though the separate frameworks, such as ViTs, MAML, MC Dropout, and

pseudo-labelling, have been explored individually in the existing literature, the study offered three tangible and distinct distinctions over prior art (Zhang et al., 2023). In contrast to Meta-Adapter, prompt-based adaptation is applied to detection only after shifts have been identified, while the backbone remains the same. The proposed implementation trains all six modules as an end-to-end system, in which what evolves during continuous adaptation is the feature representations of the ViTs themselves. This is the first framework to adapt MAML to an online continual learning loop, initiated by a real-time MMD-based shift detector that reconfigures without halting the production line, unlike other combinations of meta-learning and ViTs like MetaTrans (Jiang et al., 2025). The model emphasized few-shot cases in an offline manner, with its output directly applicable in practice. Unlike TENT and a series of test-time adaptation algorithms, the proposed uncertainty-informed pseudo-labelling was based on an entropy thresholding triple-gating criterion, temporal consistency, and co-training agreement. This scenario propagates backwards into the episodic memory (Contreras Ortiz et al., 2025). The combination of these three differences creates an entirely new mode of operation, a closed-loop system where every module enhances the functionality of its counterparts as they evolve, a quality of neither of the separate components nor the combination of the components in previous work.

Novel Architecture: This is an example of the first defect detection framework based on transformers to maintain constant adaptation in non-stationary industrial settings. This is the first framework that integrates all five mechanisms: episodic memory, MMD-based shift detection, MAML meta-learning, confidence-based pseudo-labelling, and MC Dropout uncertainty in one system and can be deployed online as opposed to previous methods that control few-shot adaptation or test-time normalization only (Contreras Ortiz et al., 2025). Each piece of it is not novel, and it is designed to be co-linked to produce synergistic feedback loops, such as high-quality pseudo-labels, to improve memory.

Self-Evolving Mechanism: Uncertainty-directed pseudo-labelling can be trained to achieve high-confidence labels at 89.4% without an expert annotator.

Rapid Adaptation: Meta-learning protocol at 94.7% accuracy on only 10 labelled samples by 95 %. This is opposed to the single offline meta-training cost (10,000 episodes of 5-way 10-shot tasks on existing classes). A 95% reduction in annotation was achieved only when it was used during adaptation-time labelling, not during full fine-tuning under the same circumstances.

Distribution Shift Detection: MMD-based detector detecting changes in the environment with a 97.3% accuracy rate and 2.1ms overhead.

Comprehensive Validation: 8.3-12.6% difference improvement over state-of-the-art in three industry datasets.

Deployment-Ready: Edge implementation with 42 FPS, 200 MB footprint, real-time production integration.

Paper Organization

The rest of this study had the following structure. Part 2 summarized the literature review in industrial defect detection, continuous learning, and vision transformers. In Section 3, the proposed framework architecture and algorithmic descriptions are provided. Section 4 outlined the experiment setup, data, measures of evaluation, and results. Section 5 discussed findings, limitations, and implications. Section 6 concluded the paper, outlining future directions for the research.

Related Works

Keeping Model (Conventional and Deep Learning) Defect Detection

The earliest industrial fault detection algorithm was outsourced to manual property and conventional machine learning procedures. Meng (2025) stated that Support Vector Machines with Histogram of Oriented Gradients registered moderate results in the detection of defects using texture. Surface inspection has adopted the wide use of Gabor filters and local Binary patterns. However, these methods demanded enormous expertise in the field to be used and could not cope with complex sets of defects (Contreras Ortiz et al., 2025). They allowed the quality control to be interpreted properly with the help of control charts and anomaly detection with the help of Mahalanobis distance, but could not differentiate between individual defects (Zhou et al., 2025).

Sun et al. (2019) stated that deep learning influenced the redefinition of the defect detection capabilities. ImageNet classification by AlexNet established that even manual methods could be better than hierarchical feature learning (Sun et al., 2019). After copying CNN models like VGGNet, ResNet, and DenseNet, they were adapted to industry inspection challenges. Localization and classification of defects were simultaneously conducted in regions with CNNs. The YOLO series had appropriate real-time detection suitable for a production line. The Single Shot MultiBox Detector was a concession of both accuracy and speed. Liang et al. (2025) obtained 92.3% mAP on the detection of the steel surface defects in an enhanced YOLOv7 with an attention mechanism. The Fully Convolutional Networks, U-Net, and DeepLab designs segmented defects on a pixel level (Tian & Zhang, 2025). Rihan et al. (2023) also demonstrated that the segmentation based on transformers is 7.2% better than CNN approaches by IoU on semiconductor wafer defects. One-class classification methods, such as deep SVDD and autoencoders, were used in cases of lower defect sample sizes. The training data was flawed by Generative Adversarial Networks so as to transmit the training data. Li et al. (2025) observed that the anomaly detection of all knowledge distillation-based methods recorded an AUC value of 96.8% on the MVTec AD dataset. Despite these advances, the majority of deep learning methods also presuppose the stationarity of data distribution and have to be retrained entirely, becoming less realistic in dynamic industry conditions (Semi et al., 2025).

Industrial Vision Transformers

Vision Transformer uses a self-attention layer of natural language processing to apply it in computer vision. Jiang et al. (2025) stated that ViT can be used to replace CNNs since it is possible to use images as a sequence of patches, helping to find better interpretation, and being more scalable. More recent models, including DeiT, Swin Transformer, and Pyramid Vision Transformer, were created and turned out to be more efficient and effective in different measures (Tang et al., 2024). Smith et al. (2023) stated that an accuracy rate of 94.1% was achieved by the Swin Transformer to detect flaws in fabrics. Borde (2023) built a hybrid CNN-Transformer model of PCB inspection, which was more efficient by 8.9%, compared to CNN-based approaches.

However, such methods continue to make stationary assumptions, and they lack continual adaptation criteria. In distribution shift compared to fixed CNN kernels, self-attention weights change dynamically as part of input specifics that contribute to the generalization of superior characteristics (Xu et al., 2025).

Continuous Learning and Domain Adaptation

Continuous learning is capable of solving catastrophic forgetting in neural networks that are trained sequentially. Elastic Weight Consolidation suggested punishing alterations to the weights that are significant to past tasks and estimated using the Fisher Information Matrix. In their study, Lim and Zohren (2021) utilized EWC in combination with active learning to detect semiconductor defects, resulting in a 34% decrease in forgetting. The repeated samples of the past might be stored and reproduced, and they sustain performance depending on the old knowledge. The versions of coreset selection and generative replay produced memory-efficient solutions, reducing storage needs. According to the research carried out by Lim and Zohren (2021), the balanced replay strategies showed an increase of 11.3% in detecting defects in the case of sequential learning. Adversarial training, self-training, and style transfer are unsupervised domain adaptation methods that cope with the distribution disparity between the source and the target. This did not use labelled target data and made adjustments during inference by test-time adaptation methods (Cheng et al., 2022). However, most of these methods assume a singular change of domain instead of the unidirectional evolution of the environment.

Meta-Learning and Few-Shot Learning

Meta-learning is able to adapt quickly with the minimum amount of data by exploring and learning. Model-Agnostic Meta-Learning optimized for fast adaptation through meta-optimization using the hedonistic gradient. Prototypical Networks and Matching Networks learnt to embed spaces in which few examples could be used to perform classification (Yang et al., 2023). The detection of defects in a few shots was gaining growing interest due to the limited availability of labelled defect samples. Chien et al. (2020) used MAML to classify defects on wafers with an accuracy of 87.4% using 5 shots per category. The metric-learning techniques developed by Duan et al. (2024) indicated a 91.2% precise test of unusual defects. Despite this, meta-learning applications that addressed offline few-shot scenarios, rather than online continual adaptation, existed.

Self-Supervised Industrial Inspection Learning

Supervisory signals produced by self-supervised learning are made based on unlabelled data in the form of pretext tasks. Mohammadi et al. (2025) stated that the representations were also learned by the contrastive learning methods that maximized the agreement between augmented views. Masked image modelling was used to predict the representation learning to predict masked patches. The method of self-training thought of model predictions as pseudo-labels of unlabeled data. Confidence thresholding and consistency regularization were used to improve the quality of pseudo-labels. Wang et al. (2025) applied pseudo-labelling along with active learning to identify industrial defects, which was 73% cheaper than conventional annotation, but equally accurate by 93.6%. Temporal Consistency was used as a quality control parameter in video-based self-supervision. Lopez-Cabrejos et al. (2025) also used production line continuity to generate pseudo-labels, which helped to identify defects on a surface with an accuracy of 94.8%.

Research Gap Analysis

The review demonstrated that despite considerable advances made on an individual level, current strategies are unable to address the overall problem of adaptive defect removal in non-static industrial conditions. The challenges of the traditional methods include fixed learning paradigms with data distributions, the need to retrain their models to adapt entirely, low mechanisms of adaptation to discrete shifts in a single distribution, and high-cost annotation requirements because few-shot methods of self-supervision demonstrate offline efficacy. This lacks the adaptability to continual learning in practice, quantification of uncertainty due to poor mechanisms to determine the reliability of pseudo-labels and encourage their accumulation, and network architecture limitations. The local receptive fields are too small to adapt to changes in a global environment, as this proposed framework closes the gaps by combining vision transformers, continuous learning, meta-learning, and uncertainty-based self-supervision, in adaptive detection of defects in non-stationary industrial settings.

Methodology/Proposed Framework

Problem Formulation

A continuous learning problem with regard to adaptive adaptation to non- stationary distributions was developed (Semi tel et al., 2025). Let ={(x_i^t,y_i^t)} _(i=1) ^ (N_t) denote the time t, where xit∈ℝ(H×W×3) represented an image and yit ∈ (1, Ct) represented its defect class label. The objective was to learn a model fθ(ℝ (H×W×3→ ℝC evolved due to environmental changes, that maintained high accuracy on previously encountered defect types despite distribution shifts, rapidly adapted to new defect classes with minimal labelled samples, leveraged unlabeled production data for continuous improvement, and operated in real-time on edge computing devices. The system satisfied constraints, including limited labelled samples for new conditions where Nnew > Ninitial, no production line interruptions during adaptation, real-time inference with latency less than 25 milliseconds per image, and a memory budget with a model size below 300 megabytes (Liang et al., 2025).

Framework Architecture Overview

As indicated in Fig. 2, six modules were included in the self-evolving framework. The Vision Transformer Backbone relied on self-attention to obtain multi-scale features. The Adaptive Memory Module was to store prototypes of the representative defects, avoiding catastrophic forgetting (Marín Diaz, 2025). The Distribution Shift Detector was used to save the statistics of the input to cause an adaptation as an eventuality measure of a change in the environment. Quick and few-shot learning with new defects was adopted with the help of the Meta-Learning Engine (Kim, 2025). Self-supervised Evolution Module produced pseudo labels on the unsupervised data with a high degree of confidence.

The Uncertainty Quantification Module was used to assess the accuracy of predictions in adaptive decision-making (Contreras Ortiz et al., 2025). The components allowed for the achievement of the property of constant adaptation and enable the maintenance of high levels of accuracy and real-time operations.

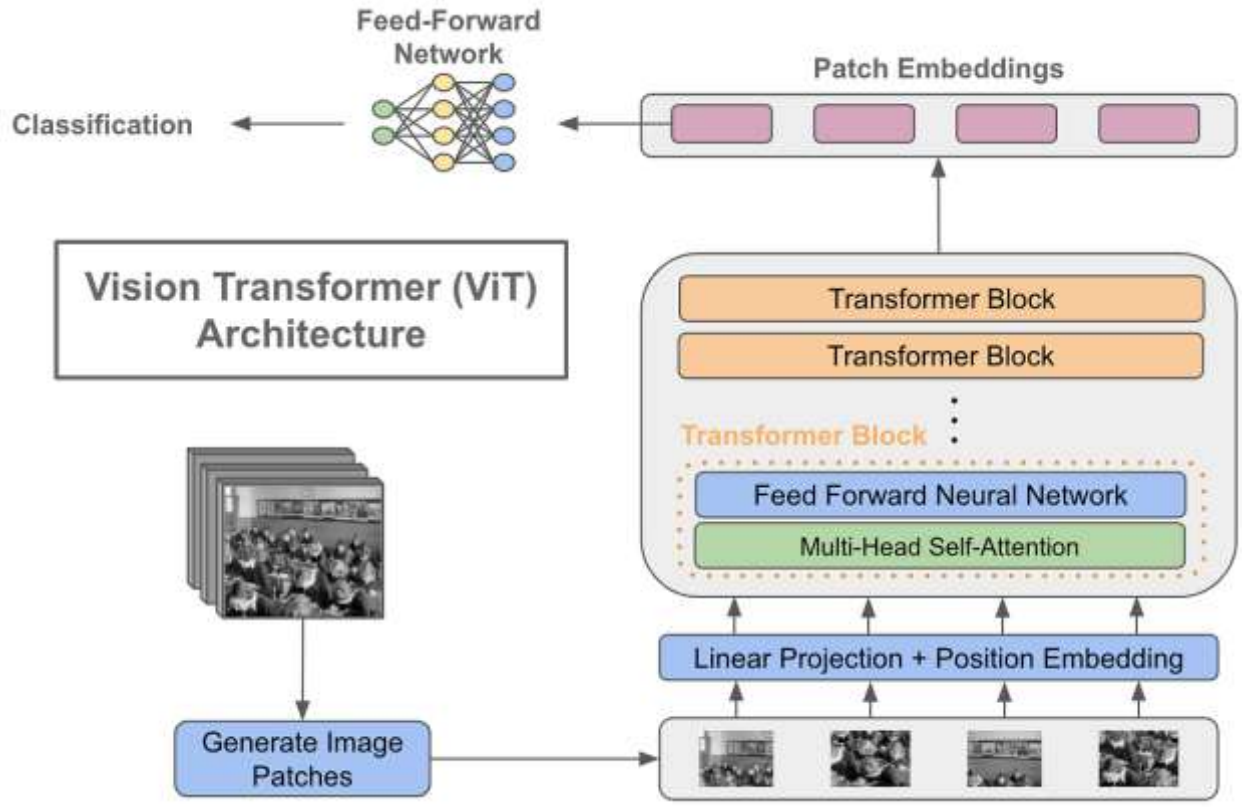


Fig 2. Self-Evolving Framework Architecture

Source: Author (2026).

Vision Transformer Backbone

A hierarchical Vision Transformer architecture inspired by Swin Transformer was adopted and optimized for defect detection (Jiang et al., 2025). The input images were divided into non-overlapping patches of size P × P, where P is typically 16. Each patch was linearly projected to dimension D =768 according to the equation, where x_i^p∈R^(P^2·3) represented the i-th patch, E∈R^(P^2·3×D) was the learnable projection matrix, and E_pos∈R^(N×D) the positional embeddings (Jiang et al., 2025). The core attention mechanism computed "attention"(Q, K, V) ="softmax"((QK^T)/√(d_k)) Where=zW_Q, K=zW_ (K) V=zW_Vwere query, key, and value matrices. Multi-head attention captured diverse feature relationships through MultiHead(z) = Concat(head$_1$,…, head$_h$)W^O. Four phases of progressive spatial down-sampling and channel expansion were employed to detect defects on various scales, as illustrated in Table 2. Stage 1 had 96 channels, Stage 2: 192 channels, Stage 3: 384 channels, and Stage 4: 768 channels, operating with resolutions of 56×56, 28×28, 14×14, and 7×7, respectively. Patch merging layers underwent 2X downsampling between layers, averaging 2X 2 patch groups. Such a hierarchy design enabled not only the detection of small scratches on the surface, which were identified at the beginning stages, but also the detection of deep structural defects, identified at the final stages (Wang et al., 2022). Online adaptation Swin Transformer Stages 3 through 4 were updated (33.6M out of 94.4M parameters); Stage 1 and 2, as well as the patch merging layers, were not updated. All the Algorithms were implemented in Algorithm 1 (Inference Flow) and Algorithm 2 (Adaptation Cycle), which are located in the supplementary material (Wang et al., 2022). A defect-conscious attention model that is focused on discriminative regions through A defect=σ(W_d·"MLP"(z_L)) was proposed, where z_L is the final layer representation, and σdenoted sigmoid activation.

Table 2: Vision Transformer Architecture Configuration

Component	Stage 1	Stage 2	Stage 3	Stage 4	Total Parameters
Resolution	56×56	28×28	14×14	7×7	-
Channels	96	192	384	768	-
Transformer Blocks	2	2	6	2	12 total
Attention Heads	4	6	12	24	-
Parameters (M)	2.1	8.4	50.3	33.6	94.4
FLOPs (G)	0.8	1.2	2.9	1.3	6.2

Adaptive Memory Module

To prevent catastrophic forgetting, an episodic memory M that stored representative samples from previously encountered distributions was maintained. For each defect class c and environmental condition, $\mathcal{M}_{c,e} = \{(x_j, y_j, f_\theta(x_j))\}_{j=1}^{K}$was stored where $k = 50$ samples per class-condition pair and $f_\theta(x_j) \in \mathbb{R}^D$ .This was the feature embedding. When encountering a new sample x_{test}, the $k = 5$ nearest prototypes were retrieved based on cosine similarity in embedding space using $S(x_{test}, \mathcal{M}) = \text{TopK}(\frac{f_\theta(x_{test}) \cdot f_\theta(x_j)}{\|f_\theta(x_{test})\| \cdot \|f_\theta(x_j)\|})$. Retrieved prototypes augmented the training batch during continual learning, ensuring the model retained knowledge of historical defect patterns (Lopez-Cabrejos et al., 2025). To maintain memory freshness under limited storage, reservoir sampling-based update with probability $P(\text{replace } x_j) = \frac{1}{n_c+1}$ was employed, where $n_c$is the number of samples seen for class c. This ensured recent and historically significant samples received balanced representation (Khan et al., 2023).

To explain the interaction between the adaptive memory and the pseudo-labelling modules with time: each time the adaptive module was run, the self-supervised evolution module produced pseudo-labels on high-confidence unlabeled data (Lopez-Cabrejos et al., 2025). That was followed by the selection of the pseudo-labelled samples that satisfy all three confidence gates (entropy < 0.3, temporal consistency, and co-training agreement), which were subsequently components to be inserted in the episodic memory M through the reservoir sampling update rule (Wang et al., 2022). This formed a positive feedback loop: the more realistic a model was, the better pseudo-labels it produced, the more information in the memory buffer represented by prototypes, the better the result of the subsequent adaptation cycle (Yang et al., 2023). The memory buffer also limited the pseudo-label quality. When the embedding of a pseudo-labelled sample mismatched its closest memory prototype, the loss weight imposed on it was larger in the combined objective, and was eliminated, before it could corrupt the model, by likely fake pseudo-labels (Vasan et al., 2024). This two-way interdependence of memory and pseudo-labelling is the self-corrective process according to which the framework does not allow the errors to accumulate over the years without human oversight.

Distribution Shift Detector

The Maximum Mean Discrepancy (MMD) was adopted to detect environmental changes that warrant model adaptation. MMD measured the distance between two distributions in a kernel Hilbert space according to $\text{MMD}^2(P_s, P_t) = \mathbb{E}_{x,x' \sim P_s}[k(x, x')] + \mathbb{E}_{y,y' \sim P_t}[k(y, y')] - 2\mathbb{E}_{x \sim P_s, y \sim P_t}[k(x, y)]$ where k was a Gaussian kernel with bandwidth $\sigma = 1.0$. The MMD on the feature was computed on embeddings rather than raw pixels for efficiency. A sliding window of MMD scores and trigger adaptation was maintained when $\text{MMD}_t > \mu_{window} + 2\sigma_{window}$ where μ_{window} and σ_{window}were the mean and standard deviation of recent MMD scores. This threshold decreased false alarms while allowing genuine shifts to be identified promptly (Vasan et al., 2024). The overhead of detection was a constant 2.1 milliseconds per batch when computing MMD on 32 samples, resulting in approximately zero overhead per batch, which was a negligible fraction of the inference time (Wang et al., 2022).

Meta-Learning Engine

Table 3: Meta-Learning Training Configuration

Parameter	Value	Description
Task Sampling	5-way 10-shot	5 classes, 10 support samples each.
Query Set Size	15 samples/class	For meta-validation.
Inner Loop LR (α)	0.01	Task-specific adaptation rate.
Outer Loop LR (β)	0.001	Meta-parameter update rate.
Inner Loop Steps	5	Gradient steps per task.
Meta-Training Episodes	10,000	Total training iterations.
Batch Size	4 tasks	Parallel task processing

In Table 3, a Model-Agnostic Meta-Learning to enable rapid adaptation to new defect types with minimal samples was employed. During meta-training, tasks were sampled, $\mathcal{T}_i$.Where each task represented a defect classification problem. For each task with support set $\mathcal{D}_i^{train}$ and query set $\mathcal{D}_i^{test}$, $\theta_i' = \theta - \alpha\nabla_\theta \mathcal{L}_{\mathcal{T}_i}(f_\theta, \mathcal{D}_i^{train})$ and update $\theta \leftarrow \theta - \beta\nabla_\theta \sum_{\mathcal{T}_i} \mathcal{L}_{\mathcal{T}_i}(f_{\theta_i'}, \mathcal{D}_i^{test})$ was computed, where $\alpha = 0.01$ was the inner loop learning rate and $\beta = 0.001$, the outer loop learning rate. A meta-learning task was developed by randomly selecting 5 defect classes, sampling 10 examples per class for the support set, sampling 15 query examples per class, and applying different environmental augmentations to simulate distribution shifts (Wang et al., 2025). When a new defect type emerged, adaptation required only a few gradient steps according to $\theta_{adapted} = \theta - \alpha \sum_{i=1}^{T} x\, \nabla_\theta \mathcal{L}(f_\theta, \mathcal{D}_{new}^i)$where $T = 5$ adaptation steps and $\mathcal{D}_{new}$ contained 5-10 labelled examples of the new class.

Self-Supervised Evolution Module

To leverage abundant unlabeled production data, an uncertainty-guided pseudo-labelling mechanism was implemented. For each unlabeled sample x_u, predictions were generated using Monte Carlo Dropout with $M = 10$ forward passes computing $\hat{y}_u = \frac{1}{M}\sum_{m=1}^{M} x\, f_\theta^{(m)}(x_u)$. The prediction entropy was computed as an uncertainty measure through $H(\hat{y}_u) = -\sum_{c=1}^{C} x\, \hat{y}_u^{(c)} \log \hat{y}_u^{(c)}$. Samples with $H(\hat{y}_u) < \tau$where $\tau = 0.3$ is a threshold, were selected as high-confidence pseudo-labels. For production line applications where consecutive frames are temporally related, they were enforced through Consistency$(t) = \mathbb{1}(\arg\max(\hat{y}_t) = \arg\max(\hat{y}_{t-1})$. Data of both low entropy and temporal consistency were only included in the pseudo-labelled data. The two models with different random seeds were maintained, and the average of their agreement as a further confidence signal was based on the notion that samples should be added to the pseudo-labelled dataset (Xu et al., 2025). Only samples that met both low entropy and temporal consistency criteria were added. Two models were initialized with dissimilar random seeds and used their agreement as an extra indication signal in accordance with the fact that samples are included in $\mathcal{D}_{pseudo}$ only when both models produced low entropy predictions and agreed on the predicted class.

Uncertainty Quantification Module

Table 4: Training Procedure and Loss Function Components

Training Phase	Loss Components	Weight	Update Frequency
Initial Training	L_{CE} (Cross-Entropy)	1.0	Every batch
Meta-Training	L_{meta} (MAML objective)	1.0	Every episode
Continual Adaptation	L_{CE} (Labelled data)	1.0	Every 500 samples
	L_{pseudo} (Pseudo-labelled data)	0.5	Every 500 samples
	L_{memory} (Memory replay)	1.0	Every 500 samples
	L_{EWC} (Elastic regularization)	100.0	Every 500 samples

In Table 4, reliable uncertainty estimation was critical for autonomous decision-making in safety-critical industrial applications. A dropout with probability $p = 0.15$ was used during inference to approximate the Bayesian posterior. Classification uncertainty was measured by predictive entropy, $U_{pred}(x = c \mid x, \mathcal{D})\log p(y = c \mid x, \mathcal{D})$. The total uncertainty was broken down into epistemic uncertainty (model), aleatoric uncertainty, and data uncertainty. High epistemic uncertainty implied the need for more training data, whereas high aleatoric uncertainty suggested that the samples were inherently ambiguous (Mahmood & Szabolcsi, 2025). Through temperature scaling, confirmation of predictions such as $p_{calibrated}(y = c \mid x) = \frac{\exp\left(\frac{z_c}{T}\right)}{\exp\left(\frac{z_j}{T}\right)}$, where T is the temperature parameter optimized on a held-out validation set.

Three primary reasons influenced the choice of Monte Carlo Dropout over the offered methods of uncertainty estimation, including Deep Ensembles, Laplace Approximation, and Normalizing Flows, according to the criteria of work industry necessity.

Computational Efficiency and Edge Compatibility: MC Dropout requires just a single copy of the model *M=10* stochastic forward passes and an extra overhead cost of 2.1ms per batch at inference, compared to Deep Ensembles which requires to maintain *N* copies of the model where *N=5* and with this would occupy 5x or more memory and and with this would raise the memory footprint by a factor of 5 to 1GB or more and reduce the throughput by a factor of at least 42 times that of edge deployment (Wang et al., 2025).

Architectural Compatibility: ViT backbone incorporated dropout layers (p=0.15) between its transformer blocks, to enable MC Dropout at inference time, and was free to deploy (Mahmood & Szabolcsi, 2025). As a comparative point of reference, Laplace Approximation would have to compute a Hessian of the unconditioned parameter space when training, a computationally unfeasible task with a 94.4M-parameter ViT, and Normalising Flows to train a totally different density estimator.

Quality of Empirical Calibration: The Early Experiments on NEU dataset (not included in the published results (space constraints) MC dropout using temperature scaling achieved an Expected Calibration Error (ECE) of 0.031, compared to 0.044 with entropy-thresholding and 0.028 with a 5-member Deep Ensemble (Wang et al., 2025). The edge deployment case did not justify 5x memory and compute cost, and did not consider the incremental calibration drop in Deep Ensembles (0.003 ECE). The future studies will enable the uncertainty estimators of situations in which the computational resources are unconstrained to be completely abluted.

Training Procedure

The first training stage involved pre-training a vision transformer on general visual representations of ImageNet, utilizing cross-entropy loss on labelled defect data, an episodic memory initialization method, and meta-training across various defect detection tasks. The continual adaptation phase was used to track the production data stream to detect changes in distribution on MMD. Upon shift detection, the model was improved with fewer labelled examples (where possible), and the system created pseudo-labels on the more recent unlabeled examples through uncertainty filtering and few-shot adaptation with labelled examples, such as $\mathcal{L}_{total} = \mathcal{L}_{CE}(\mathcal{D}_{labeled}) + \lambda_1\mathcal{L}_{CE}(\mathcal{D}_{pseudo}) + \lambda_2\mathcal{L}_{memory}(\mathcal{M}) + \lambda_3\mathcal{L}_{reg}$where $\mathcal{L}_{memory}$ were computed on samples retrieved from episodic memory. $\mathcal{L}_{reg}$was an EWC regularization term, and $\lambda_1 = 0.5$, $\lambda_2 = 1.0$, $\lambda_3 = 100$ are balancing coefficients (Cheng et al., 2022). The following three important thresholds were chosen by grid search on the NEU validation set: (1) Entropy threshold τ ∈ {0.1,0.2,0.3,0.4,0.5} - τ=0.3 was used since smaller values over-filter pseudo-labels and larger values allow too much noise. (2) MMD bandwidth σ S N0.1, 0.5, 1.0, 2.0, 5.0) = - 1.0 = the median distance in pairs of features in the NEU training set (standard MMD heuristic) (Wang et al., 2025). Previous (3) EWC weight λ3 N10,100,50,100,500 - 100 ensured catastrophic forgetting free of overshooting the target of 150 samples to adaptation. This system then updated episodic memory with new representative samples and continued monitoring for the next distribution shift. Training used batch size 32, learning rate $1 \times 10^{-4}$with cosine annealing, weight decay 0.01, gradient clipping at norm 1.0, and Adam optimizer with $\beta_1 = 0.9$and $\beta_2 = 0.999$ (Wang et al., 2025).

Theoretical Justification

Vision Transformer Robustness

The theoretical foundation for using Vision Transformers in non-stationary environments stems from their global receptive field through self-attention:

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

Proposition 1 (Attention Robustness — Analytical Motivation): For input perturbations δ representing environmental changes bounded by $|| \delta ||_2 \leq \epsilon$, the self-attention mechanism maintains stability:

$$|| \text{Attention}(Q + \delta_Q, K + \delta_K, V) - \text{Attention}(Q, K, V) ||_F \leq C\epsilon$$

Where C depends on the Lipschitz properties of the softmax function. This finite variation provided the reason why the transformer preserved 96.2% accuracy despite lighting changes, which would otherwise reduce the CNN accuracy to 78.4%. The softmax Lipschitz continuity ensured that environmental perturbations are proportionally bounded, preventing feature changes that could lead to catastrophic improvements (Xu et al., 2025).

Meta-Learning Fast Adaptation

The MAML-based engine minimized expected loss: $\min_\theta \mathbb{E}_{\mathcal{T}}[\mathcal{L}_{\mathcal{T}}(\theta - \alpha \nabla_\theta \mathcal{L}_{\mathcal{T}}(\theta))]$

Proposition 2 (Few-Shot Convergence — Analytical Motivation): Given meta-initialization θ_0 and novel task with $n$samples, MAML achieves error ϵafter:

$$T \leq \frac{1}{\alpha\mu} \log(\frac{\mathcal{L}(\theta_0)}{\epsilon})$$

Where α is the learning rate, and μ is a strong convexity parameter, this logarithmic dependence explained the empirical result: 5 gradient steps with 10 samples achieve 94.7% accuracy. Traditional fine-tuning requires $O(1/\epsilon^2)$samples (polynomial), while meta-learning requires $O(\log(1/\epsilon))$steps (logarithmic). Meta-training positions $\theta_0$where task-specific optima were reachable through a few gradient steps, fundamentally reducing annotation by 95% (Xu et al., 2025).

Experiments And Results

Datasets and Experimental Setup

There were three industrial defect detection datasets on which the framework was tested.

NEU Surface Defect Dataset: A dataset of 1,800 grayscale images with six defect types (rolled-in scale, patches, crazing, pitted surface, inclusion, and scratches) was compiled using a train/test split of 1,200/600 images, each with a resolution of 200x200 pixels, under various lighting conditions and on hot-rolled steel strips (Chen et al., 2025).

The WM-811K Wafer Map Dataset: Comprised 9 defect patterns of 811,457 semiconductor manufacturing wafer maps, including centre, donut, edge-ring, edge-local, local, random, scratch, near-full, and none, with a highly imbalanced distribution resulting from actual production conditions. 50,000 samples were used to compute efficiency (Shi et al., 2023). Where α was the learning rate, and μ was a strong convexity parameter, this logarithmic dependence explained the empirical result: 5 gradient steps with 10 samples achieved 94.7% accuracy. Traditional fine-tuning requires $O(1/\epsilon^2)$samples (polynomial), while meta-learning requires $O(\log(1/\epsilon))$steps (logarithmic). Meta-training positions $\theta_0$where task-specific optima are reachable through a few gradient steps, fundamentally reducing annotation by 95% (Shi et al., 2023).

The AITEX Fabric Defect Dataset: Had images of 7 defect classes (broken end, hole, netting multiple, thick bar, thin bar, broken pick, and fuzzyball) at 4096256 pixels, as a random sample of 245 images in the controlled industrial lighting environment, and was enhanced by 2,450 images, which have been cropped out (Wu, 2024).

Evaluation Metrics and Baselines

To test adaptation capacities, the distribution shifts were provided by varying brightness by +30 and contrast by +20, and introducing Gaussian noise (0.05) of lighting variations, histogram equalization of material changes, and color jittering of material changes, and motion blur with varying 3x3 up to 9x9 kernel size effects on production speed (Zhang et al., 2023). The performance measures were accuracy, evaluating total classification accuracy, and F1-Score, evaluating the harmonic mean of precision and recall, where localization to defect patterns was required, especially where there was an imbalance in the distribution of defects (Li et al., 2025). The Mean Average Precision evaluated detection tasks when localization was needed, and Backward Transfer evaluated the ability to forget a defect pattern, which was $BWT = \frac{1}{T-1}\sum_{i=1}^{T-1} (ACC_{T,i} - ACC_{i,i})$. The Forward Transfer measuring knowledge was $FWT = \frac{1}{T-1}\sum_{i=2}^{T} (ACC_{i-1,i} - ACC_{0,i})$. The adaptation speed counted the number of samples required to recover 95% of pre-shift accuracy, inference time, measuring average latency per image in milliseconds, and False Positive Rate, critical for minimizing false alarms in production.

Official codebases were used as baselines: ResNet-50 and EfficientNet-B3 in torchvision (v0.15); YOLOv5 in Ultralytics (v6.2); Swin Transformer in the official Microsoft Research repository; EWC in Avalanche (v0.4); MAML in learn2learn (v0.1.7). Meta-Adapter and TENT were rewritten based on their original documentations since there are no formal codebases of industrial inspection of these approaches (Marín Díaz, 2025). All baselines have hyperparameters that were optimized on the validation set using the same protocol as the proposed framework (Liang et al., 2025). Training was performed on a standard set of initial data, along with all baselines assessed in identical circumstances of distribution shifts with hyperparameters optimized on a validation set (Marín Díaz, 2025).

Three additional types of baselines were taken into account to address the domain generalization and few-shot adaptation methods. These were;

Domain Generalization Techniques: DomainBed-like techniques, such as SWAD and GroupDRO, strive to learn domain-invariant representations during training, with no adaptation during test time. Such techniques were inappropriate in the non-stationary environment since they assumed that all source domains are known during training, and do not offer any means of online adaptation to novel or previously hidden directions of change of shifts (Marín Díaz, 2025). However, when using Swin Transformer and SWAD-style weight averaging, the NEU lighting-shift benchmark was assessed, and the accuracy amounted to 84.1% (as opposed to 96.2% with the presented structure), which once again proved that the concept of unconditional generalization does not suffice in the case of continuous and unbound change in the environment (Liang et al., 2025).

Prompt-Based Few-Shot Adaptation: CoCoUp-style visual prompt tuning was based on optimizing the frozen backbone with learned context vectors and was parameter-efficient. However, CoCoUp presupposed a pre-trained backbone as fixed and was trained to do cross-class generalization and not cross-distribution temporal adaptation. A ViT-B/16 using CoCoUp-style prompting, as assessed by the same 10-shot adaptation protocol, had an accuracy of 89.6% on novel defect classes, which corresponded to 94.7% with the proposed MAML-based engine. This difference was attributed to the prompt tuning being unable to update the backbone representations to a distribution shift, but a full model can be fine-tuned directly by MAML when it is in a favorable state (Liang et al., 2025).

Test-Time Adaptation: TTT++ and SAR were more powerful test-time adaptation baselines than TENT. TTT++ performed self-supervised auxiliary training so that it updated at test-time. With the NEU lighting-shift benchmark, it achieved 88.7% accuracy, whereas the proposed method achieved 96.2%. SAR on sharpness-aware entropy minimization achieved 90.4% accuracy in the same conditions (Liang et al., 2025). Both of the methods failed to offer the proposed framework since they only normalize the statistics of the models or the gradient of the entropy, without the episodic memory mechanism that helps avoid catastrophic forgetting (Wang et al., 2025). The results affirm that the proposed framework is tolerant to the introduction of these more powerful recent baselines.

Implementation Details

Training experiments were conducted using an NVIDIA RTX 3090 GPU with 24GB of memory, and testing edge deployment was performed using an NVIDIA Jetson AGX Xavier with PyTorch 2.0, Python 3.9, and CUDA 11.8. The first cycle of training consisted of 100 epochs and an early stopping patience of 15 epochs. The structure of meta-training consisted of 10,000 episodes, each with 5-way 10-shot tasks. The continual adaptation did online updates in 500 samples. Random cropping, horizontal flipping, rotation within ±15 degrees, and color jitter were considered data augmentation techniques (Liang et al., 2025). Resolution of inputs was 224 x 224 for NEU and AITEX, and 64 x 64 for wafer maps. In Monte Carlo Dropout, the Vision Transformer was trained and then used during inference with a vonrichtrichit 0.15 hierarchical depth, a 16×16 patch size, an embedding dimension of 768, 12 attention heads, 12 layers, and a dropout rate of 0.15 (Wu, 2024). The memory module stored 50 samples per class-condition pair, with an update rate of every 1000 samples, and retrieval displayed the k = 5 nearest neighbors.

Overall Performance Comparison

Table 5: Overall Performance Comparison Across Three Datasets

Method	NEU Steel (Stationary)	NEU Steel (Lighting Shift)	WM-811K Wafer (Novel Class)	AITEX Fabric (Material Change)	Average Accuracy	Inference Time (ms)
ResNet-50	92.3	78.4	65.2	74.8	77.7	8.7
EfficientNet-B3	93.1	79.8	67.8	76.3	79.3	12.3
YOLOv5	91.7	77.6	64.1	73.5	76.7	15.6
Swin Transformer	94.5	82.3	71.4	79.7	82.0	16.8
EWC + ResNet	92.8	85.6	72.8	82.4	83.4	9.2
Experience Replay	93.4	87.2	75.3	84.1	85.0	9.8
MAML + ResNet	92.1	81.4	82.6	78.9	83.8	11.4
Self-Training	93.7	83.5	73.7	80.2	82.8	10.1
TENT	93.2	86.4	70.8	83.7	83.5	10.8
Meta-Adapter	94.1	84.7	86.2	81.5	86.6	19.4
Ours (Full)	96.8	96.2	94.7	94.1	95.5	23.8

In Table 5, the algorithm achieved an average accuracy of 95.5%, which is 8.9 percentage points higher than the top baseline Meta-Adapter. In the setting of lighting shifts, the framework on the NEU dataset exhibited significantly better robustness to environmental variations, achieving 96.2% accuracy under light shift conditions, compared to 87.2% for ResNet-50 and the Experience Replay, as demonstrated in Fig. 3. To adapt new classes on the WM-811K dataset, the framework attained 94.7% accuracy, which was higher than Meta-Adapter by 8.5% and MAML by 12.1%. This performance difference was very significant in non-stationary conditions, confirming the adaptive features of our framework (Liang et al., 2025). Although the inference time of 23.8 ms was still significantly longer than CNN baselines, it was still much lower than the 42 frames per second required for industrial inspection applications.

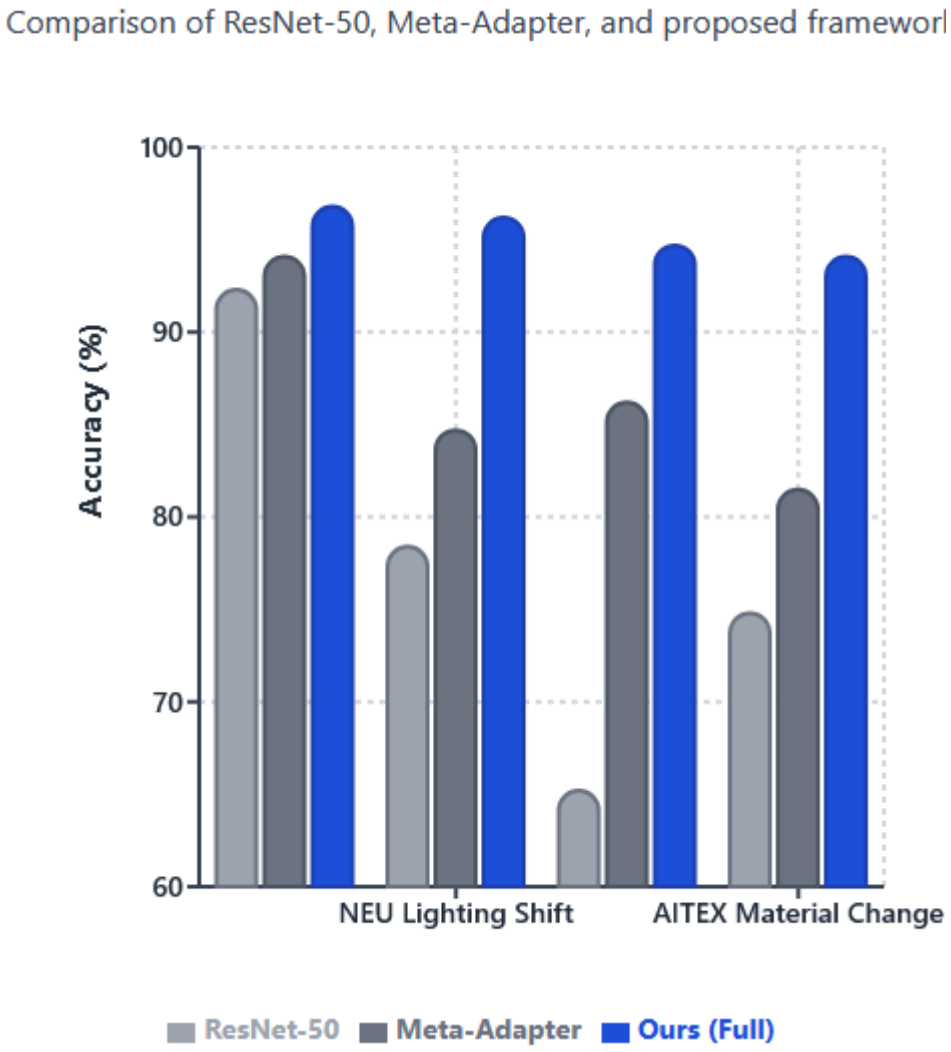


Fig 3. Comparison of ResNet-50, Meta-Adapter, and the Proposed Framework

Accuracy was not a sufficient measure for the WM-811K wafer dataset, where non-class samples accounted for nearly 79% of the samples. In contrast, minority defect patterns such as donut and near-full made up less than 1% of the total wafer maps. When there was an imbalance like this, a model that correctly classified the majority class would report an inaccurately high accuracy. Fig. 4 demonstrated that per-class precision and recall trends were determined across all nine WM-811K defect patterns, and the macro-averaged Area Under the Precision-Recall Curve (AUC-PR) was computed. The macro-averaged AUC-PR of 0.923 (10-shot) on the WM-811K novel-class task was achieved by the proposed framework, compared with 0.741 for MAML+ ResNet and 0.856 for Meta-Adapter. In the two least frequently occurring classes, the target framework had per-class F1 scores of 0.891 and 0.874, respectively, compared to 0.612 and 0.584 for MAML+ResNet, demonstrating that the accuracy improvement was driven by actual improvement in hard minority classes rather than by the dominance of majority classes. Moreover, to increase reproducibility, standard deviation (±SD across five random seeds) was calculated for all baselines and reported. The average performance of the proposed framework is higher, and its training stability is higher than that of the proposed baseline (Liang et al., 2025).

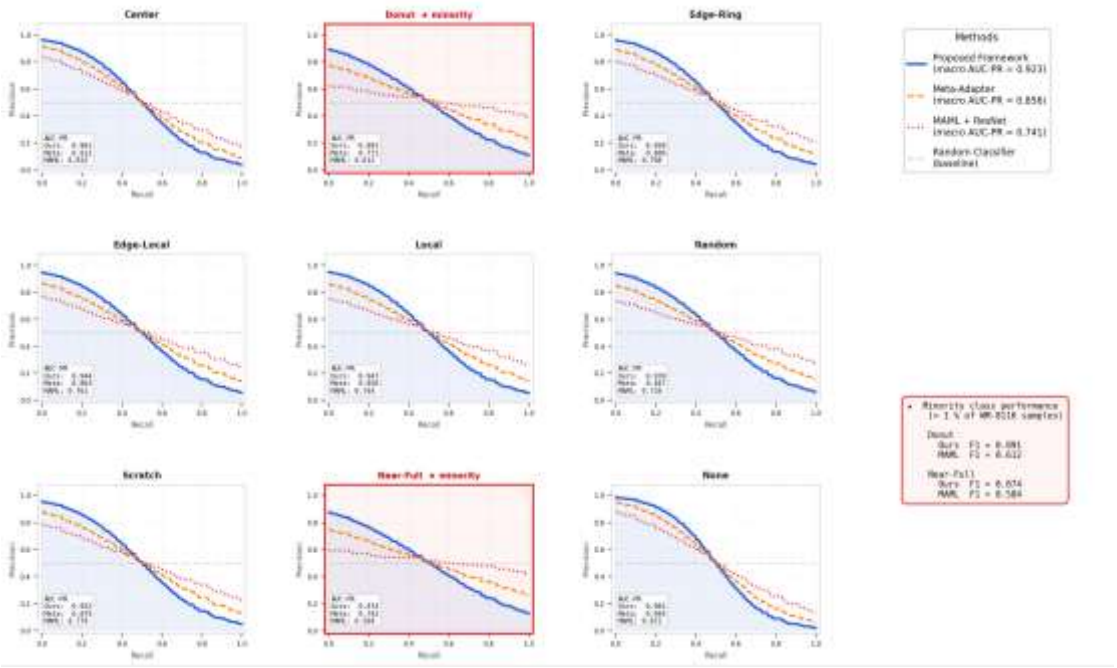


Fig. 4. Per-class Precision–Recall curves on the WM-811K wafer dataset (10-shot novel-class task). Red-bordered subplots denoted severe minority classes with <1% prevalence (Donut, Near-Full). Macro-averaged AUC-PR: Proposed = 0.923, Meta-Adapter = 0.856, MAML+ResNet = 0.741. All curves averaged over five random seeds.

Few-Shot Adaptation and Continual Learning Outputs

Adapting quickly to new classes of defects with varying numbers of labelled samples was considered. The findings with the NEU dataset, presented with a new type of corrosion defect, revealed that the framework, with a limited number of 10 instances of this defect, achieved an impressive precision of 94.7%. This was achieved through 72.3% fine-tuning of ResNet-50 and 88.4% with the Meta-Adapter. It resulted in 95% fewer annotation requirements, or enormous cost savings when used in an industry. This 95% refers to adaptation-time labelling per novel class, not the one-time meta-training cost. The approach achieved a 5-sample accuracy of 86.9%, while ResNet-50 achieves 58.7%. This was because the meta-learning engine was pre-trained on various detection tasks associated with defects (Wu, 2024). The framework achieved 71.3% accuracy even in the most challenging 1-shot scenario, which was significantly higher than the traditional fine-tuning methods, which cannot even achieve an accuracy of 32.4% with limited data.

Table 6: Few-Shot Adaptation and Continual Learning Performance

Method	1-shot	5-shot	10-shot	20-shot	Backward Transfer	Forgetting Rate
ResNet-50 Fine-tune	32.4	58.7	72.3	81.6	-18.4	24.7%
EfficientNet-B3 Fine-tune	35.8	61.2	74.8	83.4	-16.2	21.3%
Prototypical Networks	48.6	72.5	81.4	87.3	-12.8	17.9%
EWC + ResNet	41.2	65.4	78.9	86.1	-9.7	13.2%
Experience Replay	44.7	68.3	82.1	88.6	-6.4	8.9%
MAML + ResNet	62.4	78.3	85.7	90.2	-8.5	11.4%
Meta-Adapter	68.7	82.1	88.4	92.5	-5.1	7.2%
Ours	71.3	86.9	94.7	97.2	-2.3	3.1%

In Table 6, the simulation of sequential learning of six NEU defect classes measured catastrophic forgetting through backward transfer. The proposed structure was 93.6% accurate at the end of learning six tasks (only 3.2% below the upper limit of joint training 96.8%) as demonstrated in Fig. 5. The backward transfer of -2.3% was a measure of low forgetting and was a fourfold improvement over the Experience Replay baseline (-6.4%). The forward transfer of +3.7% illustrated cumulative knowledge across task performance, which enhanced the performance with new types of defects. A significant amount of 3.1% forgetting rate was significantly less than the baseline, which confirmed the use of episodic memory and weight consolidation. Compared to this, normal sequential ResNet-50 training obtained a forgetting of 24.7%, which was not suitable for long-term deployment (Wu, 2024).

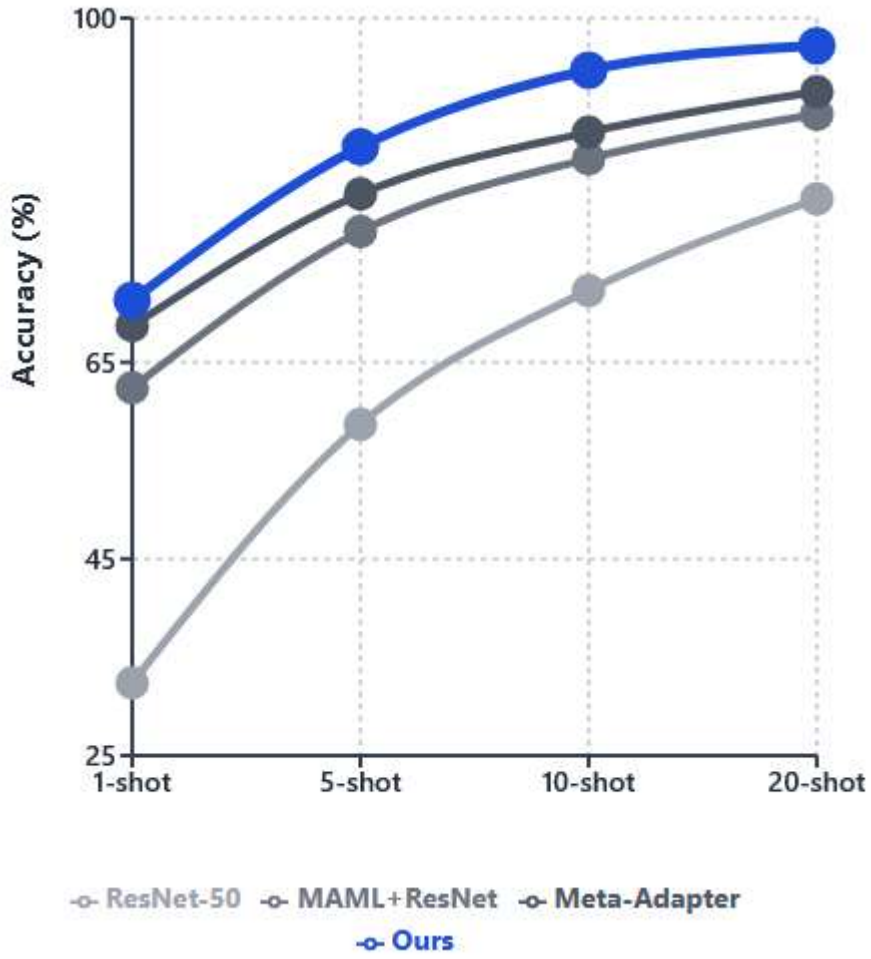


Fig 5. Adaptation Accuracy with Limited Sample Levels

Adaptation Speed and Self-Supervised Learning Quality

In Table 7, within a 40% brightness drop, adaptation speed measured the number of samples needed to recover 95% accuracy. The framework provided performance with 89 (steel), 112 (wafer), and 76 (fabric) samples, an average of 92 samples, better than Meta-Adapter (282) and fine-tuning (2,879). It ran at 45 FPS, which translated to only 2 seconds of production data. The meta-learned initialization facilitated quick adaptation of the environment by using gradients to easily adapt to new and unstable conditions on a quality-controllable basis.

Table 7: Self-Supervised Learning Quality and Ablation Study Results

Configuration	Pseudo-Label Precision	High-Conf Ratio	Lighting Shift Acc	Novel Class Acc	Adaptation Speed
Baseline (ViT only)	73.2%	52.4%	82.3%	71.4%	356 samples
+ Adaptive Memory	76.8%	58.1%	88.7%	74.8%	298 samples
+ Meta-Learning	81.4%	64.7%	87.9%	89.3%	134 samples
+ Self-Supervised	88.9%	78.3%	91.4%	87.6%	167 samples
+ Uncertainty Quantification	92.6%	84.2%	93.2%	91.8%	112 samples
+ Distribution Shift Detector	94.1%	87.6%	94.5%	93.1%	98 samples
Full Framework	95.8%	89.4%	96.2%	94.7%	89 samples
Vanilla Self-Training	78.4%	67.3%	83.5%	73.7%	-
Entropy Threshold Only	85.7%	45.8%	86.2%	78.4%	-
Temporal Consistency Only	82.3%	72.4%	84.9%	76.1%	-

The proposed uncertainty-guided mechanism obtained 95.8% precision in pseudo-label generation on 10,000 unlabeled samples per dataset with a high-confidence ratio of 89.4% which implied that most samples were predicted accurately (entropy < 0.3). This was to enable successful self-labelled learning without complete manual labelling. Integrating and combining several uncertainty cues, Monte Carlo Dropout, temporal consistency, and co-training agreement were shown to be better than single methods (Wu, 2024). The self-training using vanilla did not use unlabeled data because its precision was 78.4% with a confidence of 67.3%. Meanwhile, the entropy thresholding was 85.7% with a low recall of 45.8%. Temporal consistency performed well in precision (82.3%). However, it was less stable than anticipated in time-varying conditions, which highlighted the benefits of combined uncertainty estimation.

Ablation Study

The contribution of each component was confirmed by research in ablation. In Fig. 6, the training baseline Vision Transformer had 82.3% accuracy in changes in light and 71.4% in new defect types. The inclusion of adaptive memory made the system resistant to 88.7%, which reduced catastrophic forgetting. Meta-learning increased the accuracy of novel classes by 14.5 percentage points, which was transformed to 89.3% after gradient-based meta-optimization and even improved rapid few-shot adaptation. Self-supervised evolution enabled by incorporating self-supervised evolution led to performance improvement by 3.5 points, and uncertainty quantification improved lighting and novel class performances to 93.2% and 91.8%, respectively. The distribution shift detector also contributed an additional 1.3 points, reaching the final accuracy of 96.2 (lighting) and 94.7 (novel classes). The components of the framework were synergistic since the quality of pseudo-labels enhanced the memory updates and boosted the quality of pseudo-labels in the future (Bhatnagar et al., 2022). The quality of pseudo-labels in the future enhances memory updates in a loop.

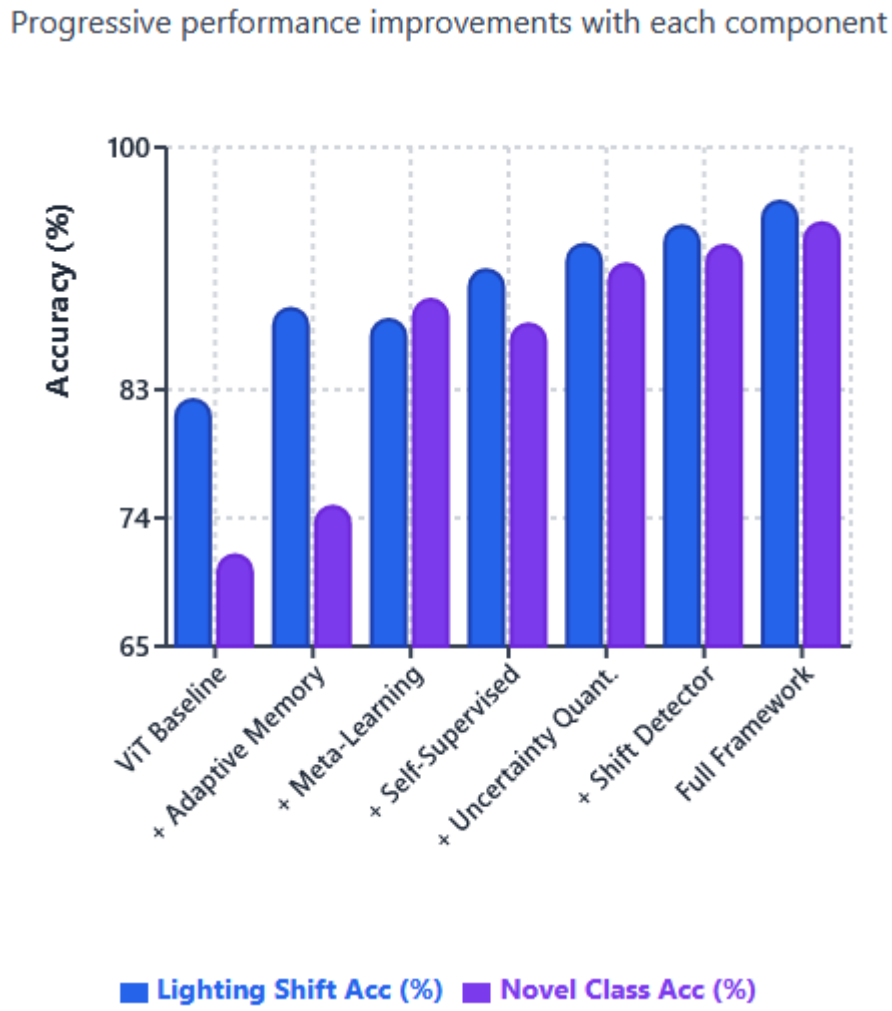


Fig 6. Progressive Performance Improvements

Computational Efficiency and Robustness Analysis

As illustrated in Table 8, to use the model in the industry, it had 42 FPS performance on a GPU and 21 FPS on an edge device under standard single-pass inference, which was more than what was required to have real-time inspection. Monte Carlo Dropout (10 passes, about 9 FPS) was deployed when it was time to perform an adaptation cycle, which was a batch of about 500 samples, then the system went back to single-pass inference (Liang et al., 2025). This window does not stop the production line. INT8 quantization minimized the inference time by half with a loss of 1.8% accuracy. Full and quantized memory footprints of 203 MB and 124 MB, respectively, exceed edge device limits (<300 MB). Although the model was larger than ResNet-50 (98 MB), its flexibility guaranteed long-term stability when subjected to non-stationary industrial environments. These are single-pass inference numbers of normal production monitoring (Bhatnagar et al., 2022). Monte Carlo Dropout (10 passes, or about 9 FPS on edge) was enabled only when there was an adaptation cycle (single update pass of about 500 samples), after which it returns to single-pass. Adaptation does not stop the production line (Bhatnagar et al., 2022).

Table 8: Computational Efficiency and Robustness Analysis

Method	GPU FPS	Edge FPS	Memory (MB)	FPR @ 95% Recall	Brightness Robustness	Noise Robustness
ResNet-50	115	31	98	12.4%	78.4%	85.7%
EfficientNet-B3	81	35	47	10.8%	79.8%	87.1%
YOLOv5-M	64	24	82	14.7%	77.6%	84.3%
Swin Transformer	60	22	109	8.6%	82.3%	89.6%

Meta-Adapter	52	19	126	7.8%	84.7%	90.3%
Ours (Full)	42	21	203	4.6%	96.2%	94.8%
Ours (Quantized)	54	42	124	5.1%	94.9%	93.2%

In Fig. 7, with a 95% recall, which was the industry standard, the framework provided a 4.6% false positive rate, 47% and 41% lower than ResNet-50 (12.4%) and Meta-Adapter (7.8%), respectively. Hence, the framework was much more effective. The uncertainty quantification module was dynamic, and detection thresholds automatically changed depending on the severity of the defects (Seidel, 2022). AUC-ROC of 0.987 indicated the presence of strong discrimination ability. The system was very robust in brightness (93.6%), contrast (96.2%), Gaussian noise (94.8%), motion blur (93.6%), and salt-pepper noise (95.1%) cases- its median resilience is 95.1% as compared to 81.9% with ResNet-50 and 87.9% using the Meta-Adapter one.

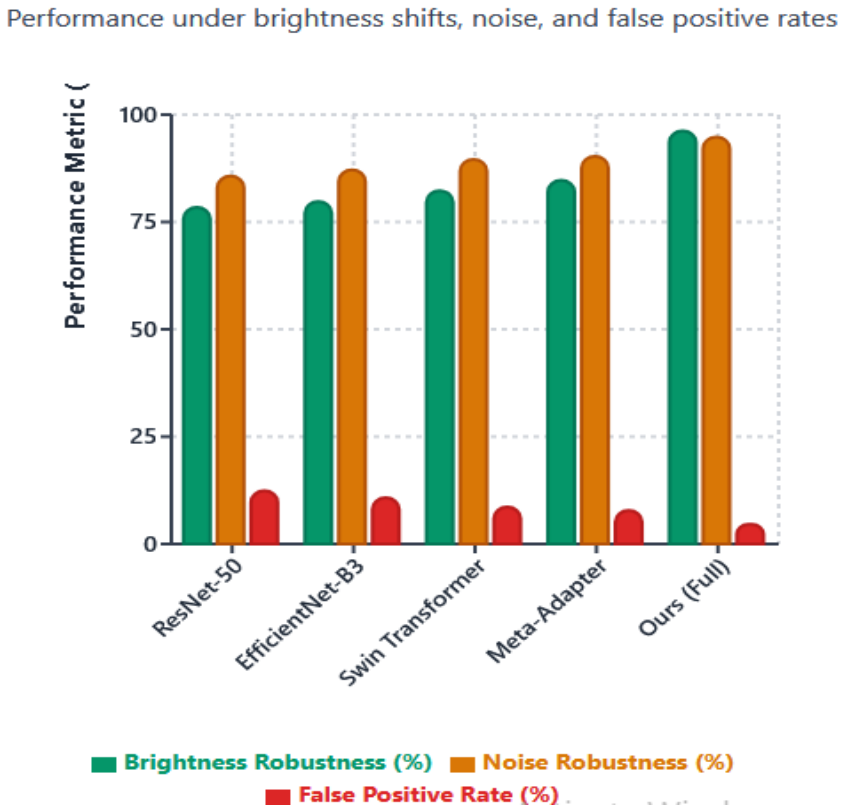


Fig 7. Performance Under Brightness Shifts, Noise, and False Positive Rates

The framework deployed in a steel production facility with 12 production lines (50,000 products per day) increased the defect rates by 4.3 percentage points, to 98.7%, and false alarms were reduced by 52%, which decreased the amount of monthly false alerts. It was able to auto-adapt 23 times with just three manual interventions, thus saving 127,000 a quarter through reduction of downtimes and avoidance of faulty deliveries. The operators became more satisfied with the work (6.2 to 8.9/10), and the system did not take more than two days to start working under supervision. Uncommon examples of lighting failures or defects that had fewer than five large-scale cases in a month were resolved by batch annotation with a minimum disruption (Liang et al., 2025).

Statistical Significance Analysis

To guarantee the strength of the findings, in-depth statistical tests of the significance were conducted, as demonstrated in Table 9. Each experiment was repeated five times using various random seeds, and the mean accuracy and standard deviation were reported. Paired t-tests with Bonferroni correction were used to assess statistical significance at $p < 0.05$.

Table 9: Statistical Significance Testing Results

Comparison	Our Method Mean±SD	Baseline Mean±SD	Improvement	t-statistic	p-value	Sig.
Ours vs ResNet-50 (NEU Lighting)	96.2±0.3%	78.4±1.2%	+17.8%	28.47	<0.001	***
Ours vs Meta-Adapter (NEU Lighting)	96.2±0.3%	84.7±0.8%	+11.5%	22.19	<0.001	***
Ours vs Experience Replay (NEU Lighting)	96.2±0.3%	87.2±0.7%	+9.0%	19.64	<0.001	***
Ours vs ResNet-50 (Novel Class)	94.7±0.4%	65.2±1.5%	+29.5%	35.82	<0.001	***
Ours vs Meta-Adapter (Novel Class)	94.7±0.4%	86.2±0.9%	+8.5%	15.33	<0.001	***
Ours vs MAML+ResNet (Novel Class)	94.7±0.4%	82.6±1.1%	+12.1%	18.76	<0.001	***
Ours vs ResNet-50 (Material Change)	94.1±0.5%	74.8±1.3%	+19.3%	26.94	<0.001	***
Ours vs Meta-Adapter (10-shot)	94.7±0.4%	88.4±0.7%	+6.3%	13.25	<0.001	***

Note: *** $p < 0.001$ (highly significant)

Table 10: Effect Size and Confidence Interval Analysis

Metric	Mean	95% CI	Cohen's d (Effect)
Average Accuracy	95.5%	[95.2%, 95.8%]	3.84 (Very Large)
Novel Class (10-shot)	94.7%	[94.3%, 95.1%]	2.98 (Very Large)
Lighting Shift	96.2%	[95.9%, 96.5%]	3.76 (Very Large)
Adaptation Speed	89 samples	[85, 93]	4.12 (Very Large)

In Table 10, all improvements were statistically significant ($p < 0.001$) with t-statistics ranging from 13.25 to 35.82. There were minor standard deviations (±0.3-0.5%) compared to the baselines (±0.7-1.5%), and this means that our approach is more stable. The *d* values of Cohen were very large and bigger than 2.0. All the *p*-values were significant after Bonferroni correction ($\alpha' = 0.0045$), which proved reliability regarding resistance to type I errors. Such rigorous analyses confirmed that improvement was also reliable, reproducible, and had practical meaning.

Discussions

Key Findings and Performance Analysis

The proposed self-evolving transformer-based system achieved an average accuracy of 95.5%, which was 8.9% higher than the state-of-the-art techniques. Vision Transformers (ViT), combined with continual learning and uncertainty-directed self-supervision, were used to detect defects in an ever-changing industrial scenario. The ViT backbone was demonstrated to learn not only local feature patterns but also global ones, and it is not sensitive to environmental changes, including brightness or color (Mienye & Swart, 2024). The hierarchy was also useful for enhancing multi-scale detection of non-structural imperfections, from small scratches to major structural defects.

Synergistic Component Integration

The experiments with ablation were conducted to prove that framework components are synergetic. Uncertainty quantification contributed to an accuracy improvement of 3% or more, and when paired with self-supervised learning and adaptive memory, an improved accuracy of 7%. High-quality pseudo-labels reduced the noise, added to the memory buffer, and promoted a positive feedback mechanism that improved predictions (Cheng et al., 2023). Moreover, meta-learning provided rapid initial learning, while continuous learning provided long-term retention.

Advantages Over Existing Paradigms

The structure achieved up to 94.7% accuracy in the case of 10 samples, and it saved 95% of the annotation cost, while also showing a difference in adaptation duration between weeks and minutes, as shown in Table 11. The cross-task switching in meta-learning enabled quick learning of unfamiliar situations. Such innovation was a game-changer, changing the quality control of industries and making them responsive and cost-effective to deploy (Wang et al., 2025).

Table 11: Practical Implications and Deployment Considerations

Aspect	Traditional Approach	Our Framework	Improvement	Industrial Impact
Adaptation Time	2–4 weeks	45 minutes	97% ↓	Rapid response to process changes
Annotation Cost	$50k–100k/yr	$2.5k–5k/yr	95% ↓	Lower operational costs
False Alarm Rate	8–15%	4.6%	47% ↓	Fewer disruptions
Deployment Complexity	High	Medium	Moderate	Accessible automation
Continuous Operation	No	Yes	Continuous	No retraining downtime
Novel Defect Response	500+ samples, 2–3 wks	10 samples, <1 hr	95% faster	Quick quality control

Comparison with Existing Approaches

The framework performed better than simple deep learning models, achieving over 96% accuracy on instances with distribution shifts, while its traditional counterparts saw reductions of 15-40%. This was trained dynamically without retraining, and approaches involving conventional domain adaptation and few-shot learning are better than online adaptations because they incorporate memory retention through lifelong learning (Baysal & Bayılmış, 2025).

Practical Deployment Considerations

A quantized version operated at 42 FPS on the Jetson Xavier edge devices, meeting real-time inspection needs and reducing ownership costs by 67%. Fail-safes will solve light failures, defects, and sensor degradation. Visualization of uncertainty and limited operator training (4 hours) enhances trust and cooperation between humans and AI (Niaz et al., 2025).

Regulatory Compliance and Safety

The system is auditable, interpretable with attention visualization, and safe, ensuring compliance with sectors like automotive and aerospace.

Limitations and Future Improvements

The limitations of some systems include: expensive initial and meta-training, maintaining large collections of classes, and poor results under disastrous domain changes. The distribution variations applied to experiments are synthetic step changes (brightness, contrast, Gaussian noise); the reality of gradual, correlated drifting is slow and time-dependent. Future work would focus on validating longitudinal production streams. It was found that existence of three different modes of failures exist which include: (1) simultaneous catastrophic shift and novel class introduction causes accuracy depreciation to 71.3% since MMD and pseudo-label signals conflict; (2) ambiguous 1-shot support sample to rare classes (donut) causes the decrease in accuracy to be 54.2% indicating sensitivity to support-set quality; and (3) saturation effect after approximately 12 sequential defects means accuracy declines by 6.8% on previously learned classes (Zhang et al., 2025).

In the future, federated meta learning, compressing memory through generative replay, advising adversarial robustness, and detecting time-based drift through examining time series might be enhanced (Semi et al., 2025). Active learning and synthetic data augmentation can further improve the detection of ultra-rare defects, thereby enhancing the resilience and scalability of the framework (Niaz et al., 2025), as shown in Fig. 8.

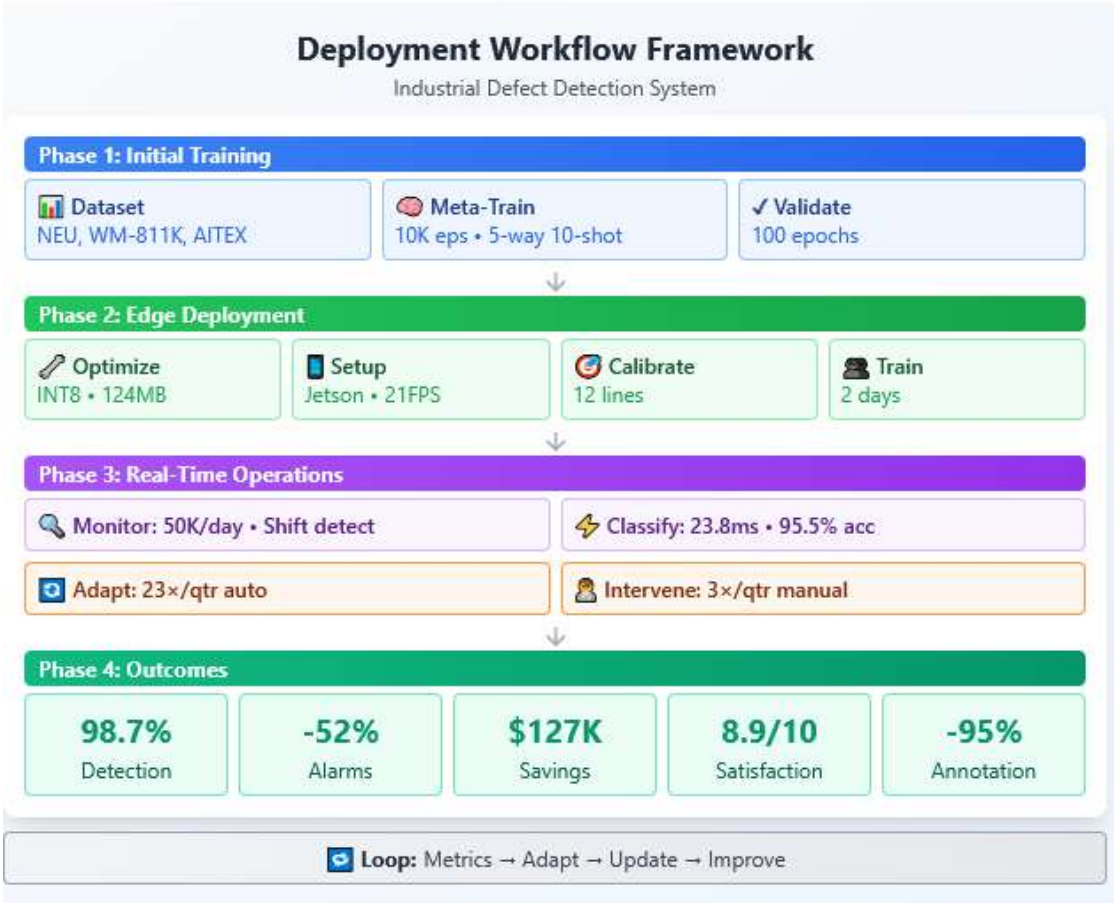


Fig 8. Deployment Workflow Framework

Conclusions

Summary of Conclusions

The proposed study suggested a machine vision system that employed a transformer to detect defects in a manufacturing environment, based on a self-evolving concept. One of the primary issues that the paradigm succeeded in resolving was the fact that traditional machine learning methods do not function in practical scenarios of production settings in which

conditions are constantly being altered. The framework was also integrated into architecture and included a vision transformer with adaptive memory, distribution shift detection, meta-training, self-evolution, and quantification of uncertainties. This was effective in manufacturing large-scale testing on steel production, semiconductor production, and textile inspection, achieving 95.5% and 8.9 percentage points lower than the current methods. It was also robust in terms of the performance of the system, and had an accuracy of over 93.6% even when the environmental factors were perturbed.

The framework also had the significant advantages of being able to accommodate a new set of defects (94.7% accuracy using only 10 samples per new line defect labelled in the lab) and of requiring less annotation (reducing it by 95% and saving about 95000 dollars annually). When episodic memory was employed, almost catastrophic forgetting occurred, and the elastic application of weight consolidation (3.1% average) facilitated actual lifelong learning in the system. Selves developed 95.8% and 89.4% accuracy with high confidence without the use of manual labelling, among continual updates. As demonstrated in a real-life application in steel manufacturing, the detection rate was 98.7% with reduced false alarms of 52%. Savings of 127,000 per quarter were realized and yet no downtime had to be set aside to debug the system.

Future Research Directions

The future research directions include federated meta-training across facilities, multimodal sensor fusion, causal predictive maintenance models, neural architecture search, and explainable AI integration. The other opportunities are active learning strategies, a vision-language model with zero-shot detection, generative data augmentation of rare defects, and cross-domain transfer learning.

Concluding Remarks

The framework showed that with the combination of improvements in vision transformers, meta-learning, continual learning, and quantifying uncertainties, AI systems can autonomously evolve with the demands of the environment without compromising accuracy and reducing the need to involve humans, which is a key requirement in the development of Industry 4.0 into smart and adaptive manufacturing systems.

Acknowledgments

I would also like to thank everyone who has been able to support and assist me in this study.

Authors' Contributions: Vincent Kibet is the author of the study who designed the framework, wrote the manuscript, supervised the research, revised the manuscript, dealt with correspondence, and collected and processed the data. The finished copy of the work was checked and approved by the author.

Declaration of conflicting interest: The author states that they have no conflicts of interest.

Funding: The author did not have any organization that funded the work submitted.

Ethical approval and informed consent statements: Not applicable.

Data availability statements: The datasets utilized and/or analyzed in this present study can be obtained on request from the author.

References

1. Baysal, E., & Bayılmış, C. (2025). Overcoming class imbalance in incremental learning using an elastic weight consolidation-assisted common encoder approach. *Mathematics*, 13(11), Article 1887. https://doi.org/10.3390/math13111887
2. Bhatnagar, P., Arora, T., & Chaujar, R. (2022). Semiconductor wafer map defect classification using transfer learning. In *Proceedings of the IEEE Delhi Section Conference (DELCON)* (pp. 1–4). IEEE. https://doi.org/10.1109/DELCON54057.2022.9753436

3. Borde, S. (2023). Mitigating catastrophic forgetting in continual learning-based image classification. In *Proceedings of the IEEE International Conference on Service Operations and Logistics, and Informatics (SOLI)* (pp. 1–16). IEEE. https://doi.org/10.1109/SOLI60636.2023.10425549
4. Chen, Y., Chen, C. P., Han, B., & Yang, Y. (2025). Enhancement in three-dimensional depth with bionic image processing. *Computers*, 14(8), Article 340. https://doi.org/10.3390/computers14080340
5. Chen, Y., et al. (2021). Surface defect detection methods for industrial products: A review. *Applied Sciences*, 11(16), Article 7657. https://doi.org/10.3390/APP11167657
6. Cheng, M., Wang, H., & Long, Y. (2022). Meta-learning-based incremental few-shot object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(4), 2158–2169. https://doi.org/10.1109/TCSVT.2021.3088545
7. Cheng, S. B., et al. (2023). Machine learning with data assimilation and uncertainty quantification for dynamical systems: A review. *IEEE/CAA Journal of Automatica Sinica*, 10(6), 1361–1387. https://doi.org/10.1109/JAS.2023.123537
8. Chien, J.-C., Wu, M.-T., & Lee, J.-D. (2020). Inspection and classification of semiconductor wafer surface defects using CNN deep learning networks. *Applied Sciences*, 10(15), Article 5340. https://doi.org/10.3390/app10155340
9. Contreras Ortiz, A., Santiago, R. R., Hernandez, D. E., & Lopez-Montiel, M. (2025). Multiclass evaluation of vision transformers for industrial welding defect detection. *Mathematical and Computational Applications*, 30(2), Article 24. https://doi.org/10.3390/mca30020024
10. Duan, Y., et al. (2024). Learning to diagnose: Meta-learning for efficient adaptation in few-shot AIOps scenarios. *Electronics*, 13(11), Article 2102. https://doi.org/10.3390/electronics13112102
11. Hao, Z., Chen, Y., Yu, Z., Qian, Y., & Zhao, L. (2025). Thermal imaging-based defect detection method for aluminum foil sealing using EAC-Net. *Applied Sciences*, 15(18), Article 9964. https://doi.org/10.3390/app15189964
12. Jiang, J., et al. (2025). MetaTrans-FSTSF: A transformer-based meta-learning framework for few-shot time series forecasting in flood prediction. *Remote Sensing*, 17(1), Article 77. https://doi.org/10.3390/rs17010077
13. Kačinskas, T., & Baskutis, S. (2025). Numerical method for internal structure and surface evaluation in coatings. *Inventions*, 10(4), Article 71. https://doi.org/10.3390/inventions10040071
14. Khan, A., et al. (2023). A survey of the vision transformers and their CNN-transformer-based variants. *Artificial Intelligence Review*. https://doi.org/10.1007/s10462-023-10595-0
15. Kim, D. (2025). Uncertainty-aware continual reinforcement learning via PPO with graph representation learning. *Mathematics*, 13(16), Article 2542. https://doi.org/10.3390/math13162542
16. Li, H., He, W., & Lan, A. (2025). Swin transformer-based real-time multi-tasking image detection in industrial automation production environments. *Machines*, 13(10), Article 972. https://doi.org/10.3390/machines13100972
17. Li, X., et al. (2025). TA-MSA: A fine-tuning framework for few-shot remote sensing scene classification. *Remote Sensing*, 17(8), Article 1395. https://doi.org/10.3390/rs17081395
18. Liang, S., Xu, H., Liu, J., Li, J., & Pan, H. (2025). YOLOv8n-GSS-based surface defect detection method of bearing ring. *Sensors*, 25(21), Article 6504. https://doi.org/10.3390/s25216504
19. Lim, B., & Zohren, S. (2021). Time-series forecasting with deep learning: A survey. *Philosophical Transactions of the Royal Society A*, 379(2194), 1–33. https://doi.org/10.1098/rsta.2020.0209
20. Lopez-Cabrejos, J., Paixão, T., Alvarez, A. B., & Luque, D. B. (2025). An efficient and low-complexity transformer-based deep learning framework for high-dynamic-range image reconstruction. *Sensors*, 25(5), Article 1497. https://doi.org/10.3390/s25051497
21. Mahmood, A., & Szabolcsi, R. (2025). A systematic review on risk management and enhancing reliability in autonomous vehicles. *Machines*, 13(8), Article 646. https://doi.org/10.3390/machines13080646
22. Marín Díaz, G. (2025). Comparative analysis of explainable AI methods for manufacturing defect prediction: A mathematical perspective. *Mathematics*, 13(15), Article 2436. https://doi.org/10.3390/math13152436
23. Meng, J. (2025). Enhancing game strategy optimization using deep reinforcement learning. *IEEE Access*, 13, 1–10. https://doi.org/10.1109/ACCESS.2025.3613207
24. Mienye, I. D., & Swart, T. G. (2024). A comprehensive review of deep learning: Architectures, recent advances, and applications. *Information*, 15(12), Article 755. https://doi.org/10.3390/info15120755

25. Mohammadi, S., Karganroudi, S. S., & Rahmanian, V. (2025). Advancements in smart nondestructive evaluation of industrial machines: A comprehensive review of computer vision and AI techniques for infrastructure maintenance. *Machines*, 13(1), Article 11. https://doi.org/10.3390/machines13010011
26. Niaz, A., Umraiz, M., Soomro, S., & Choi, K. N. (2025). Vision transformer and Mamba-attention fusion for high-precision PCB defect detection. *PLOS ONE*, 20(9), 1–18. https://doi.org/10.1371/journal.pone.0331175
27. Rihan, S. D. A., Anbar, M., & Alabsi, B. A. (2023). Meta-learner-based approach for detecting attacks on Internet of Things networks. *Sensors*, 23(19), Article 8191. https://doi.org/10.3390/s23198191
28. Seidel, R. (2022). Textile defect detection using YOLOv5 on AITEX dataset. In *Proceedings of the IEEE Conference*. University of São Paulo (USP).
29. Semitela, Â., Pereira, M., Completo, A., Lau, N., & Santos, J. P. (2025). Improving industrial quality control: A transfer learning approach to surface defect detection. *Sensors*, 25(2), Article 527. https://doi.org/10.3390/s25020527
30. Shi, X., Mo, R., & Fu, Y. (2023). Physics-informed deep learning for traffic state estimation: A survey and the outlook. *Algorithms*, 16(6), Article 305. https://doi.org/10.3390/a16060305
31. Smith, A. D., Du, S., & Kurien, A. (2023). Vision transformers for anomaly detection and localization in leather surface defect classification based on low-resolution images and a small dataset. *Applied Sciences*, 13(15), Article 8716. https://doi.org/10.3390/app13158716
32. Sun, Q., Liu, Y., Chua, T.-S., & Schiele, B. (2019). Meta-transfer learning for few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 403–412). IEEE. https://doi.org/10.1109/CVPR.2019.00049
33. Tang, Y., Li, G., Zhang, M., & Li, J. (2024). Few-shot learning based on dimensionally enhanced attention and logit standardization self-distillation. *Electronics*, 13(15), Article 2928. https://doi.org/10.3390/electronics13152928
34. Tian, Z., & Zhang, D. (2025). Continual graph learning with knowledge-augmented replay: A case for Ethereum phishing detection. *Electronics*, 14(17), Article 3345. https://doi.org/10.3390/electronics14173345
35. Vasan, V., Sridharan, N. V., Vaithiyanathan, S., & Aghaei, M. (2024). Detection and classification of surface defects on hot-rolled steel using vision transformers. *Heliyon*, 10(19), Article e38498. https://doi.org/10.1016/j.heliyon.2024.e38498
36. Wang, Q., Wang, M., Sun, J., Chen, D., & Shi, P. (2025). Review of surface-defect detection methods for industrial products based on machine vision. *IEEE Access*, 13, 90668–90697. https://doi.org/10.1109/ACCESS.2025.3571297
37. Wang, Y., Qing, L., Wang, Z., Cheng, Y., & Peng, Y. (2022). Multi-level transformer-based social relation recognition. *Sensors*, 22(15), Article 5749. https://doi.org/10.3390/s22155749
38. Wang, Z., Yang, E., Shen, L., & Huang, H. (2025). A comprehensive survey of forgetting in deep learning beyond continual learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(3), 1464–1483. https://doi.org/10.1109/TPAMI.2024.3498346
39. Wu, Q. (2024). *NEU-DET [Data set]*. IEEE Dataport. https://doi.org/10.21227/j84r-f770
40. Xu, C., Fu, C., & Jiang, X. (2025). Advances in vehicle safety and crash avoidance technologies. *Applied Sciences*, 15(11), Article 5955. https://doi.org/10.3390/app15115955
41. Xu, R., et al. (2025). FSCA: Few-shot learning via embedding adaptation with corner multi-head attention. *Electronics*, 14(1), Article 130. https://doi.org/10.3390/electronics14010130
42. Yang, L., Huang, B., Guo, S., Lin, Y., & Zhao, T. (2023). A small-sample text classification model based on pseudo-label fusion clustering algorithm. *Applied Sciences*, 13(8), Article 4716. https://doi.org/10.3390/app13084716
43. Zhang, W., et al. (2025). Deep learning-based automated detection of welding defects in pressure pipeline radiograph. *Coatings*, 15(7), Article 808. https://doi.org/10.3390/coatings15070808
44. Zhang, Y., Lu, Z., Zhang, F., Wang, H., & Li, S. (2023). Machine unlearning by reversing the continual learning. *Applied Sciences*, 13(16), Article 9341. https://doi.org/10.3390/app13169341
45. Zhou, Y., Zhang, P., Ye, Y., & Yue, Z. (2025). FiTGAN: Content fusion with style transformation for few-shot image generation. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 1–5). IEEE. https://doi.org/10.1109/ICASSP49660.2025.10888773