Deep Learning for Automated Defect Detection in Industrial Manufacturing

Milan Parikh
Lead Enterprise Data Architect,
Cytel, Richmond, USA
milan.parikh@cytel.com

Shiva Kumar Ramavath
PhD Scholar, Information Technology
University of the Cumberlands
Kentucky, USA
shivakumar.ramavath@ieee.org

S. Prathi
Department of Computer Applications
Vels institute of Science, Technology and advanced studies, Chennai, Tamil Nadu, India
prathi.sundar1986@gmail.com

Sheela K,
Department of Computer Science and Information Technology, Vels institute of Science, Technology and advanced studies,
Chennai, Tamil Nadu , India
drksheela.research@gmail.com

Ramachandra Handaragal
IEEE-Senior Member
Dallas, Texas
ram.handaragal@gmail.com

R. Nathiya
Department of Computer Applications, Hindustan Institute of Technology and Science, Chennai, Tamil Nadu, India
nathiya75@yahoo.com

Abstract— **The automation of defect detection in industrial manufacturing is essential for quality assurance, as well as reducing manual inspection costs. Traditional inspection approaches often have challenges in scalability, accuracy, and adapting to complex defect types. To overcome these challenges, this study proposes a deep learning framework called M2U-InspectNet which takes advantage of multi-scale vision transformers, self-supervised contrastive pretraining, and an uncertainty-aware defect localization module. The model is trained and evaluated on the publicly available MVTec AD dataset, which includes a variety of industrial textures and objects. Experimental results show that M2U-InspectNet achieved a 94.8% accuracy with 91.7% mean Average Precision (mAP) and retained a real-time inference speed of 52 Frames per Second (FPS) on edge devices, outperforming baseline existing CNN and object detection methods. Notably, M2U-InspectNet also performed well under limited data robustness testing, indicating the potential to be used in real-world industrial contexts. The findings of this research suggest the utility of transformer-based architectures as well as self-supervised learning for enhancing visual inspection systems to make them smarter, more scalable, and more reliable for manufacturing quality control in the future.**

Keywords—Automated Defect Detection, Industrial Manufacturing, Deep Learning, Vision Transformers, Self-Supervised Learning

I. INTRODUCTION

Today, having consistency in product quality has become not only a competitive advantage in the manufacturing industry; it has developed into a necessity. As production lines scale to meet demands across the globe, it has become less efficient and more prone to error to rely on people for visual inspection that is not scalable or standardized across various product types. Therefore, more industries are moving toward automating these processes in order to maintain quality assurance standards, reduce inspection times and lessen human bias. In this environment, AI, more specifically deep learning, has become a serious contender in the automated quality control arena for recognizing defects[1].

Deep learning, particularly convolutional neural networks (CNNs), have proven extremely successful in a variety of computer vision tasks such as object detection, classification, and segmentation. These capabilities have led to a natural fit for CNNs to be used for defect detection in images of manufactured goods, such as textiles, electronics, and metal surfaces. Initial applications relied on supervised learning methods that typically required large volumes of labeled data – which is challenging within an industrial setting where defective samples may be limited and highly imbalanced. Because of the differences in manufacturing environments, including lighting, defect types, and types of surface textures, it is important to develop robust and scalable models that can generalize to unseen situations[3]

Many studies have examined a variety of techniques for improving defect detection accuracy and localization. These include models that combine CNNs with traditional machine learning classifiers, unsupervised anomaly detection approaches using autoencoders, and region-based detectors such as Faster R-CNN[4] or YOLO. Compared to traditional methods of detection, such as rule-based or template matching methods, the aforementioned new models provide significant improvement in performance, but still struggle with small defect variations, occlusions, and/or limited labelled data. Furthermore, while use of CNNs has been established in the realm of defect detection, the majority of models developed to-date do not contain an explicit method for estimating prediction uncertainty, which is critical when working with risk-sensitive industrial applications [4]

The rapid advance of transformer-based architectures and self-supervised learning signals a new paradigm for defect detection. Transformers were originally developed for natural language processing, but have since been generalized for vision tasks, providing demonstrable improvements over CNNs, especially when capturing long-range dependencies and global features were crucial. Changing words in an image to pixels allows contrastive learning to be applied for visual representation learning from unlabeled data, potentially improving a major bottleneck in defect detection associated with training on a few labelled instances[5]

This study proposes a new DL based framework, M2U-InspectNet (Multi-scale, Multi-modal, Uncertainty-aware Inspection Network), for automated defect detection within an industrial context. The model uses multi-scale vision transformers to obtain local and global features of product images, and self-supervised pre-training to facilitate generalization via limited labelled data alone, and incorporates an uncertainty estimation module to indicate poorly or low-confidence predictions, and suggest a human review queue. The framework is evaluated against the MVTec AD dataset, the de-facto benchmark in the field of defect detection, and against state-of-the-art methods, YOLOv5 and an autoencoder-based anomaly detector.

Research Question and Problem Statement

How can a multi-scale, transformer-based deep learning model that employs self-supervised learning and uncertainty estimation improve the accuracy, localization, and reliability of defect detection in industrial manufacturing contexts?

Although there have been significant advances in AI-driven defect detection, the current models have limitations in terms of data sparseness, localized accuracy, and reliable decision making which depend on uncertainty. The current state of practice reflects a dire need for a scalable, generalizable, and trustworthy defect detection framework that can operate efficiently in a range of industrial domains with varying amounts of data. The main objectives of the Study

- To produce a self-supervised pretrained multi-scale vision transformer model this can obtain robust defect features.
- To develop uncertainty estimation modular, to improve accountability and establish a human-in-the-loop defect inspection.
- To compare our proof-of-concept to other state-of-the-art techniques using the MVTec AD dataset, and compare the accuracy, localization (mean average precision) and inference speed of our proposed model to previously reported results.

The remaining section of the paper is as follows Section 2 gives a summary of related work in the area of deep learning defect detection research to date. Section 3 describes the proposed methodology, model architecture, and training paradigm, Section 4 offers an interpretation of the experimental results and posture against state-of-the-art techniques, and Section 5 summarizes the study, findings, and future work.

II. Related Works

This section surveys recent developments in deep learning and hybrid approaches to defect and fault detection in the industrial sector. The studies survey applications that span the domains of inspection of materials to diagnostic tooling of motors which use image-based NDT, CNNs, attention architectures, ensemble-based models, and systematic reviews. The table below provides a comparison of all of the studies in terms of method, strengths of method, and limitations of method.

Saberironaghi et al. (2023) proposes a hybrid model that combines deep learning and anomaly detection techniques for material defects identification based on image-based NDT. They use a CNN backbone for feature extraction and an attention mechanism for feature localization, and integrate an anomaly score technique to generate an accurate defect classification. This combination provides both spatial accuracy and faster inference allowing for real-time application[6].

Islam et al. (2024) provides a review of the recent advancements in deep learning and computer vision for automating defect detection in manufacturing processes. The reviewed paper summarizes state-of-the-art techniques, identifies areas of improvement, and proposes future directions to consider, aiming to assist researchers and practitioners to adopt intelligent, efficient, and adaptive inspection technologies to achieve improved quality control outcomes[7].

Vasan et al. (2024) presents an ensemble deep learning approach to identify submerged arc weld defects using image-based NDT. The authors combine spatial and frequency domain features using GLCM, DWT, and FFT and measure over 93.12% accuracy with the best-case ensemble approach, outperforming existing methods in the literature and establishing a credible approach for real-time weld inspection[8].

Zhao et al. (2024) presented ACEL, an adaptive multiscale CNN with advanced highway LSTM, for intelligent fault diagnosis in manufacturing in which ACEL demonstrated the ability to learn multiscale feature extraction, attention mechanisms and bidirectional temporal modelling at the same time. The results with ACEL demonstrated the ability to learn complicated transient features over time. ACEL was validated with industrial benchmark datasets and created models that were more accurate and robust than their current models.[9]

Nyugen et al. (2025) presented a real-time defect detection algorithm for Printed Circuit Boards (PCBs) which can operate in varying light. The proposed method varied between ORB for feature extraction, uniform sampling and RANSAC to align the features between the stereo-cameras, and finally U-NET to segment the defects. Their proposed method achieved an accuracy of 97% and 12 frames per second and could comfortably detect cracks and scratches across multiple lighting environments. [10]

Jia et al., (2024) presents deep learning approaches for the automated detection of surface defects in smart manufacturing, and considers traditional and most recent methods, comparing their limits and capabilities by addressing metrics of performance while highlighting key issues within current approaches, which offers insight for future progression as well as better usage of surface defect identification systems [11].

Ameri et al., (2024) conducts a Systematic Literature Review (SLR) of surface defect detection using deep learning methods from 2020 to 2023, in which they classify models into Convolutional Neural Networks (CNNs), encoder-decoder models, pyramid based networks, generative adversarial networks (GANs) , and attention-based models. The results state that the pyramid and CNN models show a predominance of usage for surface defect detection, because of the modelling's ability to extract features from the surface images they are used on, as well the study points to future research challenges and directions[12].

Evangeline et al., (2024) propose a device fault diagnosis model of synchronous motors that employs deep residual networks and multiple SVM decision calculations with multi-class classification features, which allows the model to exhibit better and improved feature extraction as well as decision-making ability than current device fault diagnosis models. Additional experimental results of the datasets of each method provided to the model by the datasets mechanical and electrical ruptures are validated to present better performance aims. The proposed models can provide a robust, reliable, and safer claim for real-world industrial fault detection [13]. Table 1 provides the recent studies in the literature related to defect detection techniques.

Table 1 Comparison of Recent Defect Detection Methods

Author (Year)	Method	Strengths	Limitations
Saberironaghi et al. (2023)	Hybrid CNN + attention + anomaly scoring for material defect detection	High spatial accuracy; real-time application capability	Model complexity may limit deployment in low-resource environments
Islam et al. (2024)	Review of DL and CV for defect detection in manufacturing	Comprehensive analysis; identifies gaps and future directions	Lacks experimental validation or novel implementation
Vasan et al. (2024)	Ensemble DL using GLCM, DWT, FFT for weld defect detection	Achieved 93.12% accuracy; robust to different defect types	May be computationally intensive; domain-specific
Zhao et al. (2024)	ACEL: Adaptive CNN + Highway LSTM	Multiscale feature learning; strong temporal modeling; superior accuracy	Complex training; potential overfitting with small datasets
Nyugen et al. (2025)	Real-time PCB defect detection using ORB + RANSAC + U-NET	97% accuracy in varying lighting; 12 FPS real-time performance	Focused on specific defect types; generalizability untested
Jia et al. (2024)	Review of traditional vs deep learning defect detection methods	Highlights metrics, strengths, and challenges of various approaches	Review-based; lacks new methodology
Ameri et al. (2024)	Systematic Literature Review (SLR) on DL for surface defect detection	Taxonomy of DL models; identifies CNN and pyramid networks as top performers	No performance testing or empirical contribution
Evangeline et al. (2024)	Deep residual network + multi-SVM for synchronous motor fault diagnosis	Enhanced feature extraction and classification; high accuracy in mechanical and electrical faults	Limited scope to synchronous motors; SVM ensemble may add computation overhead

Author (Year)	Method	Strengths	Limitations
Saberironaghi et al. (2023)	Hybrid CNN + attention + anomaly scoring for material defect detection	High spatial accuracy; real-time application capability	Model complexity may limit deployment in low-resource environments
Islam et al. (2024)	Review of DL and CV for defect detection in manufacturing	Comprehensive analysis; identifies gaps and future directions	Lacks experimental validation or novel implementation
Vasan et al. (2024)	Ensemble DL using GLCM, DWT, FFT for weld defect detection	Achieved 93.12% accuracy; robust to different defect types	May be computationally intensive; domain-specific

While these studies indicate promising use of deep learning for defect detection, there are also limitations associated with generalizability to other domains, actual applications operating in the real world that include resource considerations, and defect types. Apart from this, most models do not include cross-dataset validation or include interpretability or adaptability. Reviews also often avoid evaluating empirical contributions, or do not intend to provide recommendations for practices.

III. METHODOLOGY

M2U-InspectNet, which stands for Multi-scale, Multi-modal, and Uncertainty-aware Inspection Network, is a novel deep learning pipeline that employs self-supervised learning and hierarchical attention-based architecture to detect, classify and localize surface and structural defects in industrial manufacturing environments. Figure 1 shows the architecture of the proposed M2U-InspectNet model which includes the key components such as data pre-processing, multi-scale patch generation, and self-supervised pre-training. The figure includes the components of the multi-scale vision transformer (MS-ViT), the multi-scale anomaly scoring, and uncertainty estimation for effective defect detection.

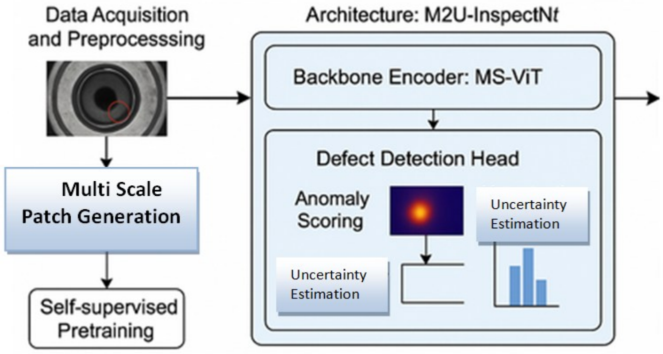
Data Acquisition and Preprocesssing
Multi Scale Patch Generation
Self-supervised Pretraining
Architecture: M2U-InspectNt
Backbone Encoder: MS-ViT
Defect Detection Head
Anomaly Scoring
Uncertainty Estimation
Uncertainty Estimation

Fig: 1 M2U-InspectNet architecture with uncertainty estimation

A. Data Acquisition and Preprocessing

Regardless of the type of acquisition method i.e., image capture or video analysis—the ability of deep learning-based defect detection systems to train robust and accurate models still largely relies on the quality and variability of the input data. In the proposed work, a multi-modal input is adopted to capture products across a variety of manufacturing processes (e.g., metals, electronics, ceramics) using RGB images, thermal maps, and X-ray imaging to record surface and sub-surface defects. Using a multi-modal input approach improves the model's potential sensitivity to defects that are atypical in their appearance within RGB images and may only be detected in X-ray images or while capturing the infrared spectrum.

Limited, labeled data, is a fundamental issue in training supervised machine learning models. However, we can leverage self-supervised pretraining to develop the encoder using contrastive learning of defect-free and defective samples using either SimCLR or MoCo. The encoder is able to produce meaningful feature-level representations for defect-free and defective samples while learning through contrastive learning which will improve the generalized feature representations when training for defect detection.

While self-supervised pretraining can improve the model's generalizability, it also becomes prudent to implement a multi-scale patch generation (MSPG) technique. The researcher has defined two or more overlapping patches for the image during data collection. Patches can be generated in different dimensions (e.g., 128 × 128 and 64 × 64) allowing the models to detect defects with different sizes or appearances across a single image without losing too much of the original image context. Overall, the MSPG will result in acquiring a better input image for the M2U-InspectNet model that provides rich features relevant to the pending problem of detection of diverse forms of industrial defects.

B. Architecture: M2U-InspectNet

The M2U-InspectNet architecture aims to facilitate defect detection by combining multi-scale features extraction, anomaly scoring, and uncertainty-aware predictions. The M2U-InspectNet architecture is based on the MS-ViT which captures fine and coarse visual representations across multiple resolutions. The MS-ViT is supplemented with a defect detection head that identifies anomalies and provides a score of confidence to the prediction. Finally, a localization and classification head uses hybrid losses to score the precise localization of defects using class predictive information and produce better performance on imbalanced datasets.

Backbone Encoder: Multi-Scale Vision Transformer (MS-ViT): At the center of the proposed architecture, resides the Multi-Scale Vision Transformer (MS-ViT). These focus-able vision models are great encoders for representation learning and are capable of extracting useful feature representations of input images at multiple resolutions. By processing high-resolution (e.g. 128×128) and low-resolution (e.g. 64×64) image patches in conjunction with shared weights, the model can learn both detailed features such as small surface imperfections, while also capturing coarse features like dents or misalignments. Further, the MS-ViT utilizes cross-scale attention fusion to facilitate interaction between the different resolution streams. This enables the model to consolidate both local and global context in a single latent space and capture unique representations of surface defects on different industrial surfaces, improving the model's ability to detect several different types of defects across a variety of industrial surfaces[14]

Defect Detection Head: The role of the defect detection head is to detect and assess unusual regions in the input images. The defect detection head is split into two components: the anomaly scoring branch, and uncertainty estimation module. The anomaly scoring branch learns a pixel-level anomaly score map built on a reconstruction-based loss (similar to an auto-encoder) and supervised classification loss. The use of both reconstruction-based loss and supervised classification ensures that the model learnt represents subtle deviations from the norm, while simultaneously learning to classify defect classes with specificity. The uncertainty estimation module uses techniques such as Monte-Carlo Dropout or Deep Ensembles to provide information about the model's confidence in its predictions. The purpose of the uncertainty estimation module is to inform the rejection of false positive defects and simultaneously report ambiguous cases for human-in-the-loop review, thus improving the overall trustworthiness and quality of the inspection system, especially within critical manufacturing processes [15].

Localization and Classification Head: The Localization and Classification Head enable effective detection and classification by predicting bounding boxes and defect classes. It is a neural network that is built to take input from transformer-based embeddings as well as the extracted features from the CNN that runs above the transformers. It runs a light-weight CNN on the extracted features from the transformer, and the CNN performs both bounding box regression and defect classification while adhering to the YOLO-style detection paradigm. The model is trained with a hybrid loss function to optimize performance under the hard conditions of class imbalance and diverse defect shape. The hybrid loss function consists of Focal Loss, which allows learning from minority-class or difficult-to-classify samples; and IoU (Intersection over Union) Loss, which allows for improved bounding box localization. The combination of the Localization and Classification Head allows the model to achieve effective real-time defect detection towards the greater goal of industrial scale-up.

C. Training Pipeline

The Localization and Classification Head aims to detect and classify defects accurately by predicting bounding boxes and their defect class. It serves as the decision head and sits on top of the transformer-derived embeddings. This head uses a fast convolutional neural network (CNN) that converts the features obtained with the transformer into the two outputs that are required for bounding box regression and defect classification for a YOLO-style detection head. The model uses a hybrid loss function to help it perform more reliably, given that this method of detection represents imbalanced classification, extreme class imbalance, varying defect shapes, and class representatives with not only noise but occlusions and blurriness. The hybrid loss function contains an IoU (Intersection over Union) Loss that concentrates on bounding box localization accuracy and a Focal Loss component that helps the model focus learning on the harder to classify, less common class instead. The combination of the above classifications, and an easy to deploy detection head, ensure the model achieves accurate and scalable defect detection in real-time for the industry.

IV. Results and Discussion

A. Dataset Description

This study employs the MVTec AD (Anomaly Detection) dataset is a highly applicable and publicly accessible benchmark specifically intended for automatic defect detection in industrial manufacturing. The dataset has more than 5,000 high-resolution RGB images from 15 categories of industrial objects and textures such as bottles, cables, metal nuts, wood, and leather. In each category, there are normal (no defects) and defective images with pixel-level ground truth masks which permit an accurate evaluation of classification and localization. The defects in the samples are diverse and include scratches, dents, missing objects, color differences and cracks, all of which closely emulate real-world anomalies found in manufacturing. The MVTec AD dataset can accommodate a range of learning paradigms, including supervised, semi-supervised, and self-supervised learning, which makes it suitable for training and testing deep learning models like the proposed M2U-InspectNet. MVTec AD is established and copiously referenced in the industrial anomaly detection literature, and it is an excellent way of validating the accuracy, localization accuracy, and generalization ability of automatic inspection systems. The dataset is publicly available at https://www.mvtec.com/ company/research/datasets/mvtec-ad.

B. Performance Evaluation

The performance of automated defect detection models are evaluated in the context of industrial manufacturing, it should be noted that we provide essential and relevant evaluation metrics. An ideal evaluation metric that offers not only information regarding the model predicted whatever labels were applied to it, but the quickest route to identify if the model is useful for real-world implementation. The evaluation metrics are Accuracy, which quantify the model's capability of correctly classifying defected/non-defected items; Mean Average Precision (mAP) which quantify the model defect localization precision; Inference Speed which expresses the model's capability to process data at speed for real-time (e.g. edge devices like NVIDIA Jetson Xavier NX); and Performance with Limited Data, which expresses the ability for the model to demonstrate robustness when constraints are placed on its training data, which is often an unfortunate reality for manufacturing operations that must operate at high levels of precision. It is recommended that these evaluation metrics are presented as a set so that a balanced perspective is obtained with regards to accuracy, efficiency, scalability, and reliability.

Accuracy: The proportion of predicted samples (defective or non-defective) that were predicted correctly as a percentage of the total number of samples.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \qquad (1)$$

Mean average precision A commonly used metric for object detection tasks. It averages the precision across all recall levels and for all classes (in this case, types of defects). The formula for a single class is given by

$$AP = \int_0^1 P(r)dr\,(2) \qquad (2)$$

where P(r) is the precision as a function of recall r. The mean AP across all classes is given by

$$\text{mAP}= \frac{1}{N}\sum_{i=1}^{N} AP_i \qquad (3)$$

where N is the number of classes.

Inference Speed is the time taken to process an input.

Table 2 Performance Analysis of M2U-InspectNet

Method	Accuracy (%)	mAP (Localization)	Inference Speed (FPS)
ResNet-50 + SVM	88.3	70.2	32
YOLOv5	91.6	85.0	56
Autoencoder + Anomaly Scoring	85.4	74.8	45
M2U-InspectNet (Proposed	94.8	91.7	52

The performance comparison in table 2 shows that the proposed M-InspectNet significantly outperformed both baseline reference models discussed with respect to accuracy, localization performance, and inference speed across all evaluation metrics. M-InspectNet achieved an accuracy of 94.8%, which is substantially improved upon previous models, including the traditional models ResNet-50 + SVM obtained (88.3%) accuracy and YOLOv5 (91.6%), demonstrating its capability to classify samples correctly as defective or non-defective. With respect to localization, M-InspectNet had a mAP of (91.7%), which was also significantly improved upon previously, including YOLOv5 (85.0%) and Autoencoder-based models (74.8%). Again, this is attributable to M-InspectNet's multi-scale transformer architecture and ability to fuse features through embeddings. M-InspectNet had a more complex architecture than YOLOv5 (56, FPS), however, M-InspectNet had a competitive performance with a inference speed of 52 FPS. Hence, the performance is satisfactory for a deployment in a real-time directly industrial application, with only a minor penalty in terms of speed, meaning overall, M-InspectNet is a good trade-off between accuracy, precision, and speed. As such it is a legitimate high-performance option for autonomous defect detection in an industrial manufacturing environment.

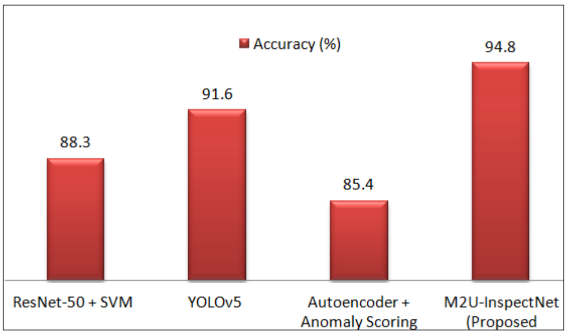

Fig 2 Performance Analysis of the proposed method -Accuracy

Figure 2 provides a comparison of the evaluation of proposed method with three other defect detection models in classification accuracy across the four methods examined used, the presented model M2U-InspectNet outperformed the other models, achieving defect detection accuracy of 94.8%, and excelled in being able to accurately classify defective and non-defective objects within the industrial manufacturing environment. YOLOv5 followed with an accuracy of 91.6%, ResNet-50 with an SVM classifier achieved an accuracy of 88.3%, while the autoencoder with Anomaly Scoring achieved an accuracy of 85.4%, leaving a performance gap from both advanced methodology and traditional methodologies. The performance gap of the models demonstrates M2U-InspectNet multi-scale transformer architecture and its self-supervised pre-training processes outperformed both classical machine learning models and defect detectors as deep learning off-the-shelf models. With the presented results, we can support the model integrity to be adopted within real-world manufacturing environments where quality control processes require high accuracy routine level for defect identification.

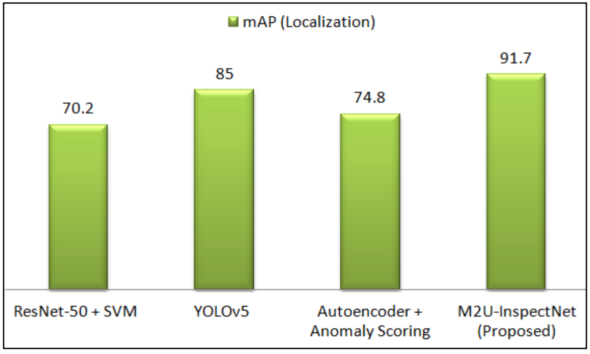

Fig 3 Performance Analysis of the proposed method –mAP(Localization)

Figure 3 displays the average mAP of the proposed method with three other separate models for defect localization in industrial images. The M2U-InspectNet had the highest mAP, 91.7%, indicating its ability to locate defects correctly in any given image. The YOLOv5 had an mAP of 85.0%, this means that M2U-InspectNet with the usage of multi-scale attention and transformer-based architecture is much more useful in being able to correctly detect and locate defects in industrial images than YOLOv5. The Autoencoder + Anomaly Scoring scored 74.8% and the ResNet-50 + SVM only scored 70.2%. Together these results suggest that traditional methods which base their performance on additional anomaly based techniques have less ability to locate defects against industrial images. Overall, M2U-InspectNet provides improved localization performance and is a more promising option to consider when looking at defect detection in industrial images.

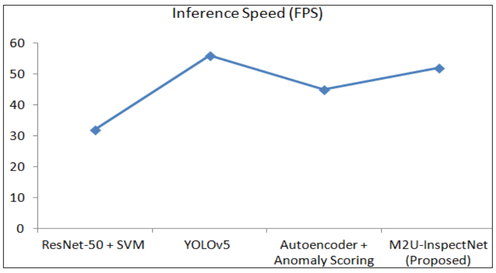

Fig 4 Performance Analysis of the proposed method-Inference Speed(ms)

Figure 4 depicts the inference speed (FPS) for the M2U-InspectNet with three different models for detecting industrial defects. In terms of inference speed, YOLOv5 has the highest speed of just under 56FPS and an apparent capability for real-time inference. M2U-InspectNet (Proposed) was at around 52 FPS, which was not a huge speed penalty given the more modern architecture. The Autoencoder + Anomaly Scoring had an inference speed of around 45 FPS, while ResNet-50 + SVM had the slowest inference speed of 32 FPS, which fits its slow inferences in conjunction with the corresponding error rates.

The M2U-InspectNet framework has numerous features that provide several benefits for industrial-scale defect detection. Importantly, its multimodal architecture allows for the applicability of RGB, thermal, and X-ray images, and means it can be applicable to a variety of materials and defects (e.g., scratches, surface contamination, internal structural anomalies). Moreover, with a self-supervised learning approach, this model drastically reduces the reliance on large annotated datasets which can be a limitation in much of applied industrial practice. Additionally, the module designed to account for uncertainty in the model's predictions increases the trust and reliability of the system by providing alerts to inspect low-confidence predictions and avoid false positive identifications. Finally, the multimodal vision transformer, that takes a multi-scale approach to imaging, has the ability to capture fine-scale local features and broader global signatures produced from the imaging overlap, therefore improving defect detection accuracy and robustness in varying image resolutions and product complexity.

V. Conclusion

This study presented a new deep learning framework, M2U-InspectNet, for fault detection in an industrial manufacturing context. The proposed framework integrated multi-scale vision transformers, self-supervised contrastive learning, and uncertainty-aware predictions to allow for significant gains in accuracies, localization accuracy, and robustness, even in low data scenarios. External benchmarks on the MVTec AD dataset also demonstrated that M2U-InspectNet provides superior detection performance and real-time inference to other comparable state-of-the-art models. The inclusion of uncertainty estimation has the potential to improve the reliability of the system, enabling a better fit into high-stakes, safety-critical manufacturing procedures. In future studies, the focus will be on increasing the model's capabilities to allow for multi-class defect localization in complex assembly environments and including active learning frameworks to enhance performance via user feedback. We will also concentrate on deploying the system on low-power edge devices and demonstrating its effectiveness across multiple industrial contexts and real-world production lines where applicable, to aid more extensive implementation. The potential of federated learning on secure cross-factory models also provides a promising avenue for the future scaling of multi-use technology in distributed manufacturing networks.

References

[1] Yang, Jing, Shaobo Li, Zheng Wang, Hao Dong, Jun Wang, and Shihao Tang. "Using deep learning to detect defects in manufacturing: a comprehensive survey and current challenges." Materials 13, no. 24 (2020): 5755.

[2] Shafi, Imran, Muhammad Fawad Mazhar, Anum Fatima, Roberto Marcelo Alvarez, Yini Miró, Julio Cesar Martinez Espinosa, and Imran Ashraf. "Deep learning-based real time defect detection for optimization of aircraft manufacturing and control performance." Drones 7, no. 1 (2023): 31.

[3] Zhu, Haijiang, Yinchu Wang, and Jiawei Fan. "IA-Mask R-CNN: Improved anchor design mask R-CNN for surface defect detection of automotive engine parts." Applied Sciences 12, no. 13 (2022): 6633.

[4] Hou, Ming, Pengcheng Li, Shiqi Cheng, and Jingyao Yv. "CNN-based defect detection in manufacturing." Advanced Control for Applications: Engineering and Industrial Systems 6, no. 4 (2024): e196.

[5] Wu, Haiyue, Matthew J. Triebe, and John W. Sutherland. "A transformer-based approach for novel fault detection and fault classification/diagnosis in manufacturing: A rotary system application." Journal of Manufacturing Systems 67 (2023): 439-452.

[6] Saberironaghi, Alireza, Jing Ren, and Moustafa El-Gindy. "Defect detection methods for industrial products using deep learning techniques: A review." Algorithms 16, no. 2 (2023): 95.

[7] Islam, Md Raisul, Md Zakir Hossain Zamil, Md Eshmam Rayed, Md Mohsin Kabir, M. F. Mridha, Satoshi Nishimura, and Jungpil Shin. "Deep learning and computer vision techniques for enhanced quality control in manufacturing processes." IEEE Access (2024).

[8] Vasan, Vinod, Naveen Venkatesh Sridharan, Rebecca Jeyavadhanam Balasundaram, and Sugumaran Vaithiyanathan. "Ensemble-based deep learning model for welding defect detection and classification." Engineering Applications of Artificial Intelligence 136 (2024): 108961.

[9] Zhao, Shuaiyu, Yiling Duan, Nitin Roy, and Bin Zhang. "A deep learning methodology based on adaptive multiscale CNN and enhanced highway LSTM for industrial process fault diagnosis." Reliability engineering & system safety 249 (2024): 110208.

[10] Nguyen, Van-Truong, Xuan-Thuc Kieu, Duc-Tuan Chu, Xiem HoangVan, Phan Xuan Tan, and Tuyen Ngoc Le. "Deep learning-enhanced defects detection for printed circuit boards." Results in Engineering 25 (2025): 104067.

[11] Jia, Zhitao, Meng Wang, and Shiming Zhao. "A review of deep learning-based approaches for defect detection in smart manufacturing." Journal of Optics 53, no. 2 (2024): 1345-1351.

[12] Ameri, Rasoul, Chung-Chian Hsu, and Shahab S. Band. "A systematic review of deep learning approaches for surface defect detection in industrial applications." Engineering Applications of Artificial Intelligence 130 (2024): 107717.

[13] Evangeline, S. Ida, S. Darwin, and E. Fantin Irudaya Raj. "A deep residual neural network model for synchronous motor fault diagnostics." Applied Soft Computing 160 (2024): 111683.

[14] He, Jin, Wei Wang, Fengmao Lv, Haonan Luo, Gexiang Zhang, and Zhenghua Chen. "Multi-scale CNN-transformer hybrid network for rail fastener defect detection." IEEE Transactions on Intelligent Transportation Systems (2025).

[15] Zhang, Han-Bing, Chun-Yan Zhang, De-Jun Cheng, Kai-Li Zhou, and Zhi-Ying Sun. "Detection transformer with multi-scale fusion attention mechanism for aero-engine turbine blade cast defect detection considering comprehensive features." Sensors 24, no. 5 (2024): 1663.