IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 16 December 2025, accepted 16 January 2026, date of publication 22 January 2026, date of current version 2 February 2026.

Digital Object Identifier 10.1109/ACCESS.2026.3657141

RESEARCH ARTICLE

Diffusion-Augmented Industrial Anomaly Detection Leveraging Attention-Empowered Multi-CNN Transformer Fusion

MD MAHINUR ALAM[1], **ABDULAZIZ YAGOUB BARNAWI**[2], **(Member, IEEE), MOHTASIN GOLAM**[3], **KANITA JERIN TANHA**[3], **AND TAESOO JUN**[3], **(Member, IEEE)**

[1]Department of Electrical and Computer Engineering, Georgia Southern University, Statesboro, GA 30458, USA
[2]Department of Computer Engineering, IRC for Intelligent Secure Systems, King Fahd University of Petroleum and Minerals, Dhahran 31261, Saudi Arabia
[3]Department of IT Convergence Engineering, Kumoh National Institute of Technology, Gumi 39177, Republic of Korea

Corresponding author: Taesoo Jun (taesoo.jun@kumoh.ac.kr)

This work was partly supported by the Institute of Information and Communications Technology Planning and Evaluation (IITP)–Innovative Human Resource Development for Local Intellectualization program grant funded by the Korea government (MSIT) (IITP-2025-RS-2020-II201612, 40%), the ITRC (Information Technology Research Center) grant funded by the Korea government (MSIT) (IITP-2025-RS-2024-00438430, 30%), the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2018R1A6A1A03024003, 30%), and the Regional Innovation System and Education (RISE) (Regional Growth Innovation LAB) program through the Gyeongbuk RISE Center, funded by the Ministry of Education (MOE) and the Gyeongsangbuk-do, Republic of Korea (2025-rise-15-105).

ABSTRACT Deep learning models have demonstrated remarkable performance in anomaly detection tasks, particularly when large datasets with sufficient anomalous samples are available. However, in industrial settings, collecting an adequate number of defective images for training remains a significant challenge. Additionally, the complexity and diversity of anomaly patterns in industrial environments often make it challenging for a single model to capture the entire range of features effectively. To address these challenges, this paper proposes the Versatile Attention Re-weighted Stable Diffusion (VARSDiffusion), a diffusion-based model for generating realistic defective samples. Furthermore, it introduces an Attention-Empowered Multi-CNN Transformer-Fused model architecture to extract diverse feature sets from anomaly images. A transformer-based meta-learner is employed to combine these features, enabling the model to detect both subtle and large-scale anomalies more effectively. The proposed approach addresses issues such as class imbalance, poor image variability, and the risk of overfitting by incorporating extensive data augmentation techniques, including the application of Gaussian noise, to improve model generalization. The model was evaluated on the MVTec LOCO AD and MVTec AD datasets, which present real-world challenges such as sparse data and imbalanced classes. The results demonstrate the robustness and reliability of the model, achieving AUROC scores of 98.15% and 97.56%, respectively, outperforming existing state-of-the-art approaches. This novel framework provides a robust and adaptable solution for industrial anomaly detection, even in environments with limited or imbalanced data.

INDEX TERMS Anomaly detection, multi-CNN, stable diffusion, transformer, meta learner, manufacturing system.

I. INTRODUCTION

Maintaining product quality in industrial production in the 21st century is essential due to the ever-increasing demand for production and the continuously rising standards for quality [1]. Inspecting production defects is essential for ensuring the quality of manufactured goods. By identifying and addressing flaws, it can enhance the overall quality of the products and maintain high standards for customer satisfaction. In modern production environments, even minor anomalies such as scratches, dents, or misassembled components can lead to significant quality failures, safety hazards, and costly recalls [2]. Detecting any anomaly becomes essential for quality or cost control in production and critical to the integrity of industrial processes [3]. As manufacturing standards and output rates continue to rise, manual inspection becomes impractical and error-prone.

The associate editor coordinating the review of this manuscript and approving it for publication was Mehul S. Raval.

A vast impact in anomaly detection research has been achieved using computer vision and deep learning [4], [5]. This has opened the possibility of developing high-level models and systems where anomaly detection might become autonomous [6]. Different models and strategies have been proposed to leverage the effective capabilities of CNNs in handling the complexities of production defect detection. These deep learning models have gained popularity due to their remarkable performance in image detection [7]. Despite the advancements made, the current state-of-the-art anomaly detection models still face significant limitations when dealing with industrial anomalies. These anomalies exhibit considerable variation across different production environments, making them difficult to capture through conventional methods. Furthermore, these models often fail to generate detailed feature representations, which are crucial for identifying subtle and complex anomalies that might otherwise go undetected [8]. The heterogeneous nature of industrial environments, coupled with the diverse characteristics of anomalies, presents unique challenges that many traditional models are unable to overcome effectively [9]. As a result, there is a growing need to develop more advanced and tailored approaches that can better address these challenges [10]. Additionally, the limited nature of labeled datasets increases the challenges associated with model generalization and overfitting [11]. There are problems with employing complex and fuzzy anomaly detection systems in industrial settings [12]. These include uneven class distribution, limited data diversity, and the requirement for models to demonstrate resilience and efficacy in various operating contexts [13]. Therefore, developing novel and unique strategies that enhance the performance and generalization abilities of deep learning models is necessary to tackle these issues [14], [15].

This research proposed an anomaly detection framework for complex industrial environments, utilizing an Attention-Empowered Multi-CNN Transformer-Fused model in convergence with Versatile Attention Re-weighted Stable Diffusion. The model integrates attention modules with multiple CNN networks to capture both local and global features, enabling efficient anomaly detection by focusing on relevant data regions [16]. A transformer network is used to fuse the features extracted from CNNs, enhancing the model's ability to identify intricate anomaly patterns by understanding contextual dependencies across different feature sets. The Stable Diffusion model generates synthetic anomalies, addressing class imbalance, while Versatile Attention Re-weighting further improves the quality of these generated images by focusing on the most significant anomaly features. This combined approach forms a robust framework capable of tackling the challenges of industrial anomaly detection. More precisely, the notable contributions of this work could be outlined as follows:

1) This paper proposed a novel framework for industrial anomaly detection aimed at enhancing production excellence within manufacturing systems. The framework integrates a Stable Diffusion model for generating synthetic defective images and a multi-CNN transformer-fused model, which efficiently identifies anomalous products along the production line.
2) The proposed model enhanced the feature extraction capabilities of CNN architectures by integrating a novel Attention mechanism. Additionally, it developed MCTNet, a multi-CNN fusion model that combines multiple CNN models using a Transformer as a meta-learner. This approach enables the model to focus on critical features and adapt quickly to variations in anomalies, achieving more accurate and efficient defect detection compared to existing fusion-based methods.
3) To address real-world industrial dataset limitations like limited availability of defective images and class imbalance, the proposed model employed a Versatile Attention Re-weighted Stable Diffusion (VARSDiffusion) method to generate realistic defective samples. Furthermore, advanced Data Augmentations were incorporated to enhance the model's generalization and robustness, along with the addition of Gaussian noise, thereby improving its performance on unseen data.
4) A comprehensive performance comparison of the proposed method is demonstrated and conducted against existing methods on the MVTec LOCO AD and MVTec AD datasets. This comparison serves to validate the reliability and efficacy of our proposed method in the context of anomaly detection.

The structure of the paper is organized as follows: **Section** II describes the related work in this area. **Section** III explains the system and proposed model architecture, along with the components, optimization, and training techniques used in this approach. **Section** IV depicts the testbed configuration. **Section** V summarizes the empirical findings and assesses the method's effectiveness. **Section** VI concludes the study and presents potential avenues for future research in anomaly detection for manufacturing systems.

II. RELATED WORK

Recent advancements in anomaly detection for manufacturing systems have leveraged diverse methodologies, ranging from traditional machine learning to cutting-edge generative models; yet, critical challenges persist in achieving robust, scalable, and generalizable solutions. This section provides a comprehensive synthesis of contemporary approaches across multiple paradigms, highlighting their contributions and fundamental limitations to contextualize our proposed framework.

A. UNSUPERVISED METHODS

Unsupervised techniques have gained substantial traction due to their ability to detect anomalies without the need for extensive labeled datasets. Cohen and Hoshen [17] proposed SPADE, a deep pyramid correspondence-based method that localizes anomalies through multi-scale feature matching. While SPADE demonstrates strong performance on benchmark datasets, achieving competitive AUROC scores, it exhibits critical vulnerabilities: alignment dependency on fixed reference images renders it susceptible to dataset variability, the multi-resolution feature pyramid architecture incurs significant computational overhead (up to 2.5x inference time compared to single-scale methods), and limited adaptability to novel defect types not represented in the reference set. Tao et al. [18] explored unsupervised defect detection via dual-Siamese networks, addressing scenarios with limited labeled anomalies.

Recent developments have introduced more sophisticated unsupervised approaches. Roth et al. [19] developed PatchCore, achieving state-of-the-art performance with 99.6% AUROC on MVTec AD through patch-based memory banks and nearest-neighbor matching. However, PatchCore faces scalability challenges, as its memory requirements grow quadratically with dataset size, making it impractical for large-scale industrial deployment. Additionally, the method struggles with fine-grained texture anomalies and exhibits inconsistent performance across different material types.

Xing and Li [20] addressed pixel-level anomaly detection through a Partition Memory Bank (PMB) module, introducing spatial partitioning to improve localization accuracy. Despite improvements in computational efficiency, their approach faces limitations in capturing the semantic contextual information crucial for complex industrial scenarios. The reliance on skip connections for low-level feature integration introduces ambiguity in model interpretability, necessitating further empirical validation across heterogeneous defect types commonly encountered in manufacturing environments.

B. SEMI-SUPERVISED AND FLOW-BASED APPROACHES

Semi-supervised methods attempt to bridge the gap between unsupervised techniques and fully supervised learning. Rudolph et al.'s DifferNet [21] leverages normalizing flows for defect identification, demonstrating robustness in limited-data scenarios through probabilistic modeling. The framework shows promise in handling distribution shifts common in industrial settings. However, critical limitations emerge: struggles with high-dimensional image data due to inherent constraints of normalizing flows, pixel-wise localization proves suboptimal for defects exhibiting spatial irregularity such as texture variations or structural misalignments, and computational complexity increases exponentially with image resolution.

Recent flow-based approaches [22], [23] have sought to address these limitations by employing hierarchical modeling and enhanced architectures. However, fundamental constraints in capturing complex, multi-modal defect distributions persist, highlighting the need for more adaptable architectures capable of handling diverse anomaly morphologies encountered in real-world manufacturing scenarios.

C. TRANSFORMER-BASED APPROACHES FOR INDUSTRIAL DEFECT DETECTION

The introduction of Vision Transformers and attention mechanisms has brought significant advances to industrial anomaly detection, addressing limitations of purely convolutional approaches in capturing global dependencies and long-range spatial relationships crucial for defect identification.

Zhang et al. [24] proposed a detection transformer with multi-scale fusion attention for aero-engine turbine blade defect detection, achieving 96.8% accuracy through attention-based channel-adaptive weighting mechanisms that effectively handle imbalanced datasets. However, their method focuses primarily on single-domain applications and lacks the generalizability required for diverse industrial scenarios with varying defect characteristics. Huang et al. [25] introduced a sparse cross-transformer network for surface defect detection, incorporating residual layer modules while reducing computational complexity by 40% compared to full attention. Despite maintaining detection accuracy, their approach struggles with strict real-time industrial latency requirements (sub-100ms inference time).

Li et al. [26] explored transformer-based meta-learning for bearing fault diagnosis, demonstrating 15% improvement over traditional CNN methods in few-shot scenarios through adaptive attention mechanisms and gradient-based meta-learning. However, the framework faces limitations in handling diverse anomaly morphologies across different industrial domains and requires domain-specific fine-tuning for optimal performance.

These transformer-based approaches highlight the potential of attention mechanisms in industrial defect detection but reveal critical gaps in unified frameworks that can effectively handle multiple CNN architectures while maintaining computational efficiency suitable for industrial deployment constraints.

D. DIFFUSION-BASED ANOMALY GENERATION

Recent breakthroughs in diffusion models have opened up new avenues for addressing data scarcity in industrial anomaly detection through the generation of high-quality synthetic defects, representing a paradigm shift from traditional data augmentation techniques.

Zuo et al. [27] introduced a training-free anomaly generation framework (AAG) utilizing stable diffusion without requiring additional training data. Their cross-attention enhancement (CAE) and self-attention enhancement (SAE) mechanisms generate realistic anomalies while maintaining industrial context coherence, achieving Inception Scores of 1.94 and FID scores below 45 on the MVTec AD dataset. However, their method lacks the versatile attention re-weighting mechanisms necessary for precise anomaly mask alignment and defect morphology control. Sun et al. [28] proposed AnomalyAny, leveraging stable diffusion for diverse unseen anomaly generation across multiple industrial domains with superior LPIPS scores indicating high

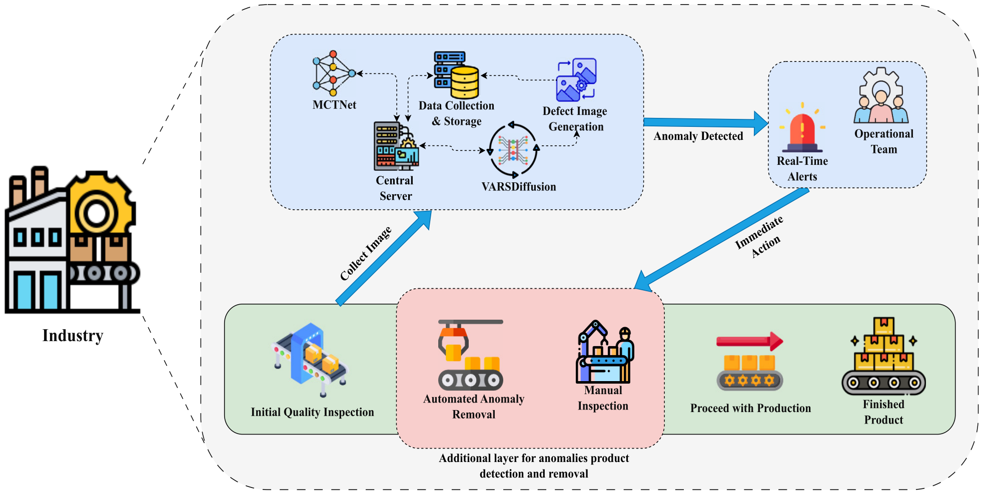

FIGURE 1. Proposed industrial anomaly detection framework with centralized data processing for defect generation using VARSDiffusion and MCTNet for anomaly detection.

perceptual similarity to real defects. Despite promising dataset diversity expansion, their approach faces limitations in controlling specific defect characteristics required for precise morphology control and adherence to physical formation constraints. Dai et al. [29] explored SeaS for few-shot industrial anomaly generation, achieving 12% improvement over traditional GAN-based methods in scenarios with extremely limited defective samples (fewer than 10 per class). However, the framework struggles with consistency across different defect types and exhibits mode collapse issues when generating complex, multi-component defects, which are common in assembled products.

These diffusion-based methods demonstrate significant potential for addressing the scarcity of industrial data. However, existing approaches lack sophisticated attention re-weighting mechanisms that can adaptively focus on critical anomaly features while ensuring precise defect mask alignment, particularly in complex industrial environments with diverse anomaly patterns and strict quality control requirements.

E. ADVANCED META-LEARNING AND ENSEMBLE APPROACHES

Recent developments in meta-learning have shown promise for industrial anomaly detection, particularly in scenarios with limited labeled data and high variability in defect characteristics. However, current approaches face significant challenges in effectively combining multiple feature extraction architectures and adapting to novel defect types.

The integration of meta-learning with transformer architectures has gained attention for few-shot defect detection scenarios. Lee et al. [30] proposed XDNet, a cross-domain meta-learning framework that leverages squeeze-excitation modules, anti-aliasing filters, and contrastive learning for visual inspection tasks. Their approach combines ensemble feature extractors with non-parametric classifiers, achieving approximately 17% improvement over existing state-of-the-art methods on the MVTec AD dataset reformulated for n-way k-shot classification. However, their method shows domain sensitivity, performing well on textured materials (leather: 84% normalized accuracy) but struggling with complex object domains (capsule: 22% normalized accuracy), and faces computational challenges when scaling to larger industrial deployments.

Recent ensemble approaches have evolved beyond simple majority voting to address the limitations of uniform weighting. Zhou et al. [31] proposed multimodal fusion combining visual, thermal, and vibration data, achieving 94.7% accuracy with dynamic weighting that adapts to sensor reliability and environmental conditions. However, the approach requires extensive sensor infrastructure and faces real-time processing challenges due to the complexity of multi-modal data fusion. Liu et al. [32] introduced selective ensemble methods addressing data imbalance and computational efficiency through intelligent model selection, achieving comparable performance to full ensemble methods while reducing computational overhead by 60%. Despite these advances, existing ensemble methods often employ static architectural combinations that disregard complementary strengths of different CNN backbones and lack sophisticated fusion mechanisms for optimal feature integration.

These developments highlight the critical need for sophisticated fusion mechanisms that dynamically weight contributions from different models based on reliability, complementary strengths, and performance consistency, particularly when combining diverse CNN architectures with transformer-based meta-learners in dynamic industrial environments.

F. RESEARCH GAPS

Current literature analysis reveals four critical limitations that significantly hinder practical industrial anomaly detection deployment:

First, insufficient integration of advanced generative models with attention mechanisms. While diffusion-based approaches [27], [28], [29] demonstrate strong generation capabilities, they lack sophisticated attention re-weighting mechanisms for precise anomaly localization and feature prioritization in complex industrial contexts. This limitation results in generated samples that may lack precise defect characteristics required for effective model training, particularly for subtle anomalies in high-precision manufacturing environments.

Second, limited fusion of diverse CNN architectures with transformer meta-learners. Existing methods [24], [25], [26] typically focus on single architecture optimization or simple ensemble approaches without effectively leveraging complementary strengths of multiple CNN backbones through intelligent transformer-based fusion. This results in suboptimal feature utilization and limits the system's ability to handle diverse defect types that different architectural paradigms may better capture.

Third, inadequate handling of real-world industrial constraints. Current approaches [30], [31] often assume clean, well-aligned data and struggle with practical challenges, including severe class imbalance (often 1:100 normal-to-defect ratios), limited data variability, diverse anomaly morphologies, and strict latency requirements (sub-100ms inference) commonly encountered in manufacturing environments.

Fourth, lack of unified frameworks combining state-of-the-art generation with sophisticated fusion mechanisms. While recent works address either anomaly generation [27], [28] or multi-architecture fusion [24], [26] independently, no existing approach effectively integrates both paradigms within a single, computationally efficient framework.

To address these critical limitations, this work introduces a unified framework combining Versatile Attention Re-weighted Stable Diffusion (VARSDiffusion) for controllable defect generation with MCTNet, a transformer-based meta-learner that intelligently fuses attention-enhanced CNN architectures (VGG16, DenseNet121, ResNet50). Unlike existing approaches employing static fusion mechanisms, our method employs adaptive attention re-weighting for both generation and fusion processes, enabling superior performance across diverse industrial scenarios while maintaining computational efficiency suitable for real-time deployment. The framework addresses data scarcity through sophisticated synthetic generation, handles architectural diversity through intelligent fusion, and maintains practical deployment viability through optimized computational design.

III. METHODOLOGY

A. ANOMALY DETECTION SYSTEM

In the Manufacturing System (MS), the product line comprises various applications and modules designed to address multiple aspects of the manufacturing process. This system is characterized by its flexibility and scalability, allowing for the optimal selection and configuration of modules tailored to specific production requirements. The MS aims to enhance operational efficiency, reduce costs, improve product quality, and enable data-driven decision-making to optimize manufacturing processes. A critical component of this is the continuous monitoring and detection of anomalies within the production line to ensure quality and operational excellence. The proposed framework integrates a Stable Diffusion model to generate synthetic images, thereby enhancing the capabilities of anomaly detection. As illustrated in Fig. 1, the process begins with the collection of images and data during the initial quality inspection stage. These images are sent to a central server, where they are analyzed using an AI-based anomaly detection model [33]. To enhance the robustness of defect identification, the proposed VARSDiffusion model generates additional synthetic defect images, thereby enriching the training dataset for the anomaly detection model MCTNet, which enables more accurate and advanced defect identification. Upon detecting anomalies, the system triggers real-time alerts for immediate action by the operational team [34]. Defective products are directed to an automated anomaly removal process to address identified defects efficiently. Once the automated process resolves the anomalies, the products undergo manual inspection to ensure that no defects are overlooked and that quality standards are met [35]. After this verification step, the products re-enter the production line and ultimately reach the finished product stage.

B. VERSATILE ATTENTION RE-WEIGHTED STABLE DIFFUSION

In this approach, the VARSDiffusion model generates synthetic images that address data scarcity and imbalance issues commonly encountered in industrial defect detection. By stochastically producing a broader and more diverse set of defect images, the approach reduces the dataset acquisition cycle, expands low-volume datasets, and mitigates overfitting. The focus is not on replicating exact real-world parts, but rather on exploiting the model's inherent stochasticity and capacity to generalize, thereby improving the robustness and utility of the training data. As shown in Fig. 2, the process begins by encoding the original image into a lower-dimensional latent space using a trained encoder E, producing a latent vector z_0. Within this latent space, the forward diffusion process progressively adds noise at each timestep t, and can be represented as:

$$z_t = \sqrt{\alpha_t} \cdot z_0 + \sqrt{1 - \alpha_t} \cdot \epsilon \tag{1}$$

Here, α_t denotes the timestep-dependent noise-scheduling coefficient, which varies with t to control how much of the original signal is preserved at each diffusion step. Both $\sqrt{\alpha_t}\, z_0$ and $\sqrt{1 - \alpha_t}\, \epsilon$ represent standard scalar vector multiplications, with $\epsilon \sim \mathcal{N}(0, I)$ denoting Gaussian noise. The square-root terms follow the standard diffusion parameterization, where the signal and noise components are scaled according to their respective standard deviations. The

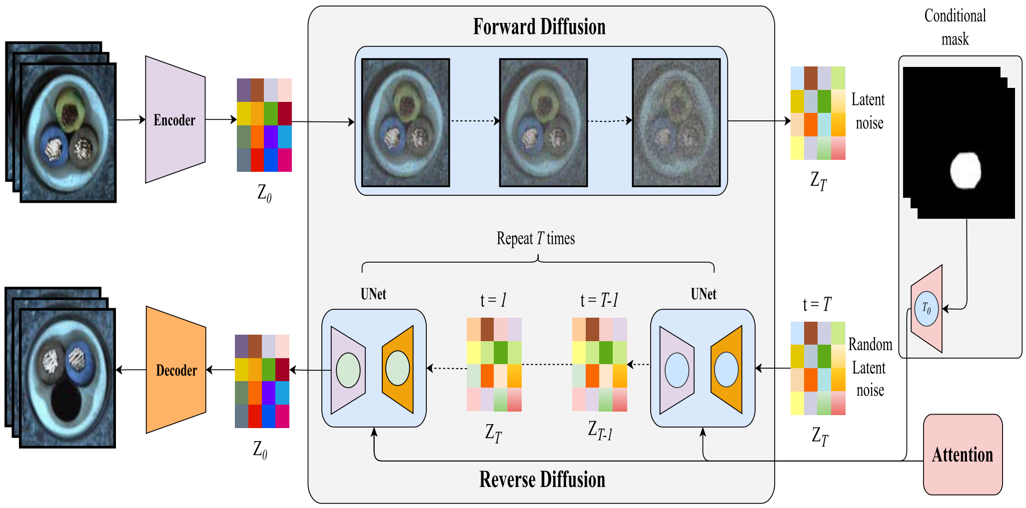

FIGURE 2. Overview of proposed VARSDiffusion model architecture for industrial defective sample generation.

probability of generating the full sample trajectory $p_\theta(z_{0:T})$ is governed by the model parameters θ, with $p(z_T)$ serving as the prior distribution. The transition probabilities between steps follow:

$$p_\theta(z_{t-1} \mid z_t) = N\left(z_{t-1}; \mu_\theta(z_t, t), \Sigma_\theta(z_t, t)\right) \tag{2}$$

where, in the reverse diffusion process, z_{t-1} is generated from the current latent state z_t via the learned transition $\mu_\theta(z_t, t)$, so each step depends only on the previous state and the current diffusion timestep t in a Markovian manner. $\mu_\theta(z_t, t)$ and $\Sigma_\theta(z_t, t)$ are the timestep-dependent mean and covariance derived from the model. The coefficient α_t, introduced in the forward noising process, specifies the noise level at timestep t and is used to encode t inside $\mu_\theta(z_t, t)$. Consequently, μ_θ is implicitly a function of α_t. To optimize these parameters, a latent diffusion loss $\mathcal{L}_{\text{LDM}}$ is applied:

$$\mathcal{L}\text{LDM} = \mathbb{E}z, \epsilon \sim N(0, 1), t\left[|\epsilon - \epsilon_\theta(z_t, t)|_2^2\right] \tag{3}$$

where the subscript of the expectation indicates that the loss is averaged over samples of (z, ϵ, t). Here, z is drawn from the latent data distribution, $\epsilon \sim \mathcal{N}(0, 1)$ denotes standard Gaussian noise, and t is the diffusion timestep. $\epsilon_\theta(z_t, t)$ denotes the model's denoised prediction at timestep t, and minimizing this loss aligns the reconstructed latent representation with the noiseless target. Following T denoising iterations guided by the U-Net-based diffusion backbone, a Versatile Attention Re-weighting mechanism is incorporated to refine feature prioritization. Specifically, at each denoising step t, a weight map w_m is computed based on the pixel-level differences between the generated image $\hat{x}_0$ and the normal sample y within the anomaly mask m:

$$w_m = |m|_1 \cdot \text{Softmax}\left(\frac{1}{|m \odot (y - \hat{x}_0)|_2^2}\right) \tag{4}$$

This weight map w_m is used to modulate the cross-attention mechanism, directing more focus to regions with less pronounced anomalies:

$$\text{RW-Attn}(Q, K, V) = \left(\text{Softmax}\left(\frac{QK^\top}{\sqrt{d}}\right) \odot w_m\right) \cdot V \tag{5}$$

where Q, K, and V are the query, key, and value matrices derived from the latent code z_t and spatial anomaly embedding e, and d is the dimensionality of the key vectors. This adaptive attention re-weighting enhances the alignment between generated anomalies and anomaly masks by emphasizing areas that require more accurate anomaly generation.

The refined latent vector is then decoded by a decoder D to produce the final restored image. This VARSDiffusion enhancement technique effectively manipulates feature shapes, pixel intensities, and contextual details. By enriching positional, environmental, and contrast-related information within the generated images and applying adaptive attention re-weighting for critical feature focus, this approach yields diverse, high-quality synthetic samples. Such outputs foster more robust and efficient model training in industrial defect detection tasks.

C. PROPOSED MULTI-CNN TRANSFORMER FUSION

This study proposes the MCTNet model, which integrates the feature outputs of multiple pre-trained deep-learning models, each enhanced with a specialized attention mechanism, to enhance anomaly detection performance. This method focuses on three highly efficient architectures: VGG16, DenseNet121, and ResNet50, chosen for their complementary feature extraction capabilities and superior performance in anomaly detection compared to other existing state-of-the-art models, as further explained and demonstrated in the evaluation section. The VGG16 is equipped with a

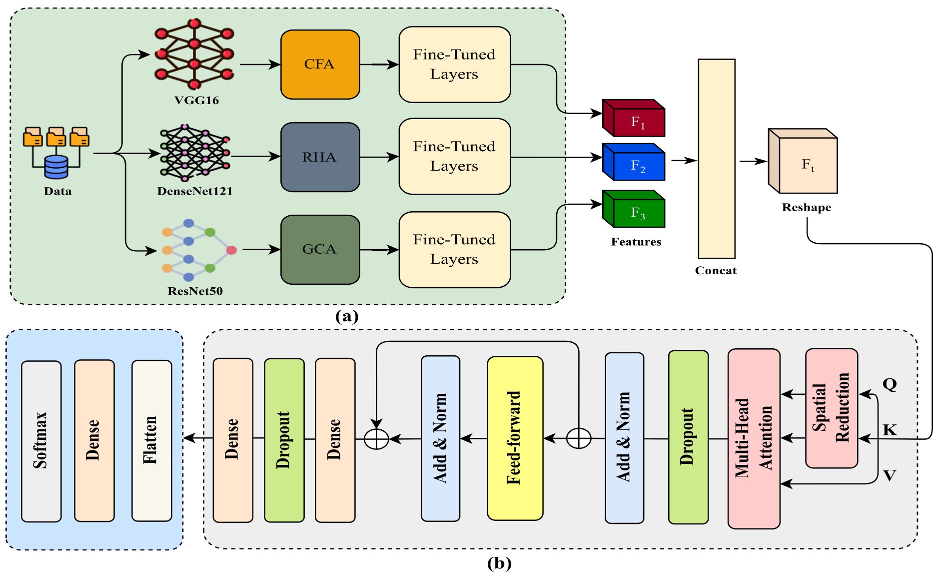

FIGURE 3. Proposed Attention-Empowered Multi-CNN Transformer-fused Model (a) Illustrates the feature aggregation of the Multi-CNN model along with the Attentions and (b) Shows the Meta Learner fusion Transformer.

Channel Focus Attention (CFA) to emphasize informative feature channels. Additionally, DenseNet121 is combined with a Regional Highlight Attention (RHA) to highlight critical spatial regions, and ResNet50 is augmented with a Global Context Attention (GCA) to capture long-range dependencies. The details of this attention mechanism process are explained in the forthcoming sections. By utilizing these distinct, attention-empowered architectures, the system obtains three refined feature representations, denoted as F_1, F_2, F_3, where $F_i = \text{Model}_i(X)$ and X represents the input data. These features are subsequently concatenated:

$$F_t = [F_1; F_2; F_3] \tag{6}$$

where $F_t \in \mathbb{R}^{N\times d}$, N is the number of samples, and d is the dimensionality of the concatenated feature vector. After fine-tuning, the three backbone attention branches produce fixed-length feature vectors: the VGG16–CFA branch yields a 512-dimensional vector, the DenseNet121–RHA branch yields a 1024-dimensional vector, and the ResNet50–GCA branch yields a 2048-dimensional vector. These vectors, denoted $F_1 \in \mathbb{R}^{512}$, $F_2 \in \mathbb{R}^{1024}$, and $F_3 \in \mathbb{R}^{2048}$, are concatenated to form the fused feature representation $F_t \in \mathbb{R}^{3584}$ that is passed to the transformer module, as illustrated in Fig. 3, the transformer dynamically identifies the most salient features and exploits interdependencies across the combined representation. The core operation within the transformer is the multi-head self-attention mechanism, which projects F_t into a query (Q), key (K), and value (V) subspaces:

$$Q = F_t W_Q, \quad K = F_t W_K, \quad V = F_t W_V \tag{7}$$

where $W_Q, W_K, W_V \in \mathbb{R}^{d\times d_k}$ are learned weight matrices, and d_k is the projection dimensionality. The scaled dot-product attention is then computed as:

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right) V \tag{8}$$

To concurrently attend to multiple representation subspaces, multi-head attention is employed:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W_O \tag{9}$$

where $W_O \in \mathbb{R}^{h\cdot d_k\times d}$ is the output projection matrix and h is the number of heads. Additionally, a spatial reduction mechanism is integrated within the transformer to refine and prioritize essential spatial features, further enhancing the quality of the fused representation [38], [39]. Following the attention layer, a feed-forward network (FFN) introduces non-linear transformations:

$$\text{FFN}(x) = \max(0, xW_1 + b_1)W_2 + b_2 \tag{10}$$

where W_1, W_2 and b_1, b_2 are learnable parameters. Residual connections and layer normalization are applied to stabilize training and improve generalization:

$$x_{\text{norm}} = \text{LayerNorm}(x + \text{output}) \tag{11}$$

After multiple transformer layers have refined the fused feature representation, it is reshaped and fed into fully connected layers, culminating in a softmax layer:

$$\hat{y} = \text{Softmax}(F_{\text{output}} W_c + b_c) \tag{12}$$

where W_c and b_c are the weights and biases of the final layer. This advanced framework utilizes attention-driven models

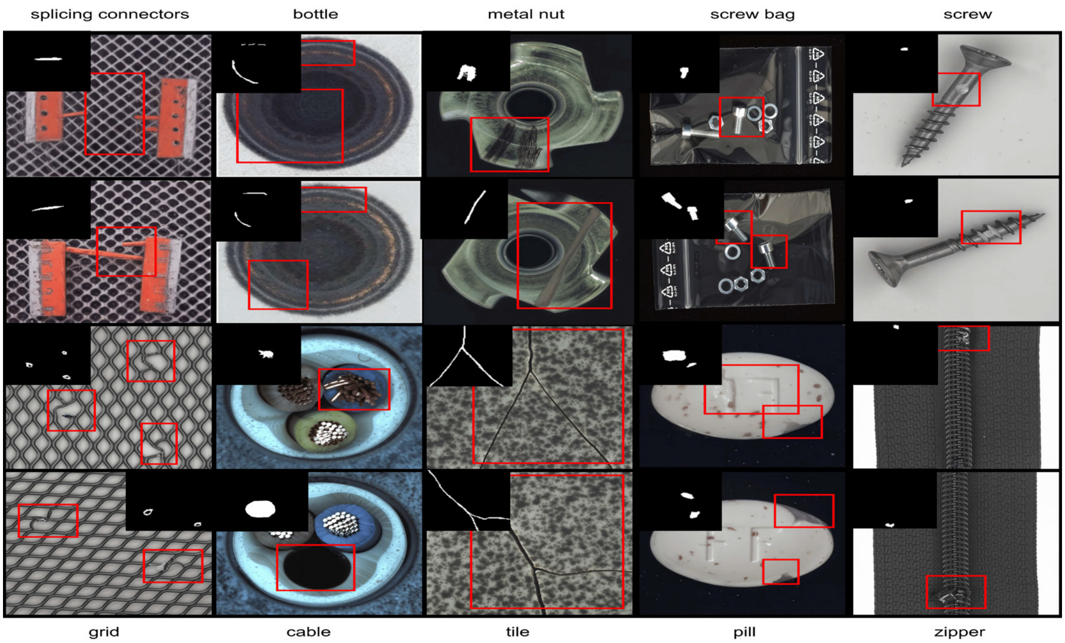

FIGURE 4. Illustration of the generated high-quality defective samples, highlighting the anomaly areas in red based on the corresponding masks. The samples represent industrial product images from the MVTec LOCO AD [36] and MVTec AD [37] datasets.

combined with transformer techniques to emphasize the most critical features. It reduces unnecessary information and improves the way features work together. This leads to enhanced anomaly detection, making it more reliable, accurate, and capable of handling different datasets effectively.

D. TRANSFORMER AS META LEARNER

The Transformer model serves as a meta-learner, enabling the system to adapt to new anomaly detection tasks rapidly with minimal retraining. In our implementation, the transformer module is realized as a single transformer encoder layer comprising one multi-head self-attention block and a subsequent position-wise feed-forward sublayer, both equipped with residual connections and layer normalization. Here, the core functionality of the Transformer as a meta-learner lies in its self-attention mechanism, which allows it to focus on the most relevant features across different industrial defects detected by various models. By leveraging the relationships between features learned from multiple anomaly detection models, the Transformer can generalize across new defects. During training, the gradient-based optimization technique Model-Agnostic Meta-Learning (MAML) is used to optimize the parameters θ of the Transformer. MAML enables the model to learn a shared initialization θ^* that can be quickly adapted to new defects with minimal data. For each new defect T, initial learning rate α, and anomaly detection loss $\mathcal{L}_T$, the model performs a gradient update to adapt the parameters:

$$\theta^{(T)} = \theta - \alpha \nabla_\theta \mathcal{L}_T(\mathcal{M}_\theta(T)) \tag{13}$$

This enables the Transformer to efficiently fine-tune its parameters and adapt to new defects, allowing the system to detect previously unseen anomalies with high accuracy and improve performance across diverse industrial scenarios.

E. DEFECT SYNTHESIS AND DATA AUGMENTATION

VARSDiffusion addresses the challenge of insufficient data samples in minority classes by generating synthetic defective images, as shown in Fig. 4. The process begins with a noise vector $z \sim \mathcal{N}(0, I)$ which is iteratively refined using a reverse diffusion process guided by a defective mask M. This mask M, extracted from existing samples, ensures that the generated images retain realistic defect patterns. Formally, the denoising step at each iteration t can be expressed as:

$$x_{t-1} = \text{Denoise}(x_t, M, \theta) \tag{14}$$

where x_t represents the noisy image at step t, θ denotes the model parameters, and Denoise is the learned reverse diffusion function. By progressively applying this process, VARSDiffusion enriches the dataset with diverse and high-quality defective images, enhancing the model's capacity to detect anomalies effectively. This is followed by data augmentation using TensorFlow libraries. Each image underwent a series of random transformations: images were randomly rotated within the range of [−360°, 360°], and each color channel was randomly shifted within ±30 (approximately ±15% of the 8-bit color range [0, 255]). To account for variations in illumination, pixel intensity was randomly altered within the range of [−20, 20], affecting all color channels uniformly. Additionally, pixel intensity was randomly scaled within the range of [0.9, 1.5]. Random perspective transformations were applied to simulate a variety of distances and viewing angles. Images were scaled vertically and horizontally within

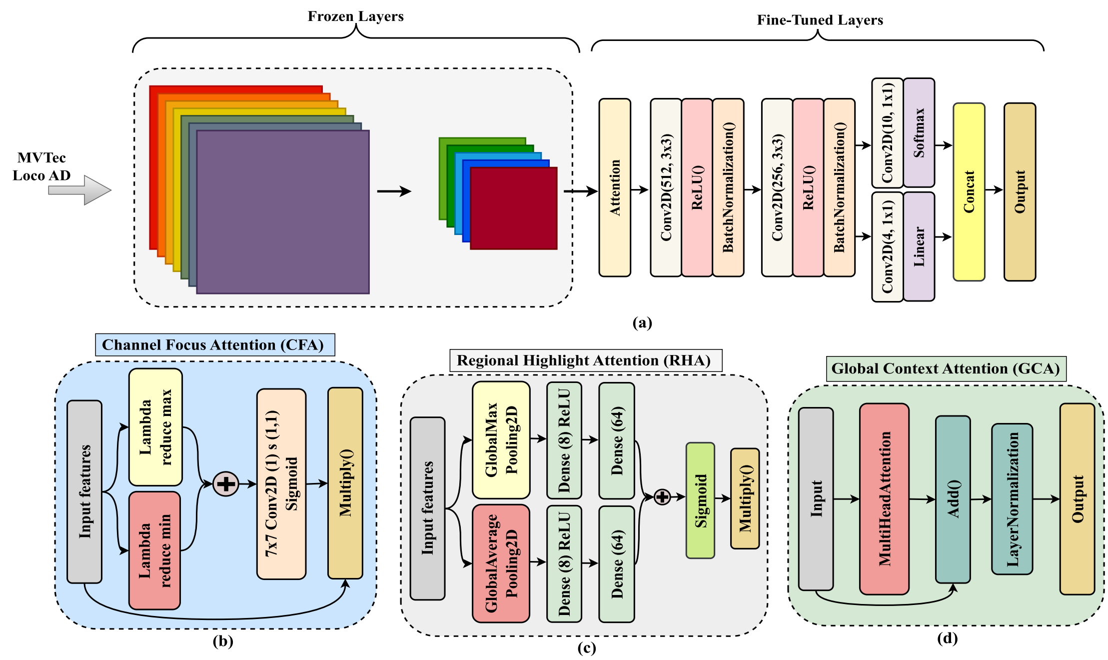

FIGURE 5. **Illustration of the attention-enhancement and fine-tuning process for the Multi-CNN model architecture; (a) depicts the overall fine-tuning of the Multi-CNN models, while (b), (c), and (d) represent the corresponding attention mechanisms designed for efficient feature extraction.**

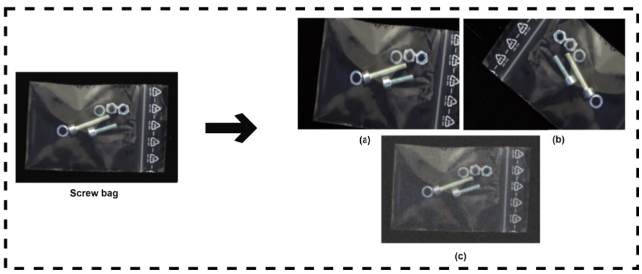

FIGURE 6. **Examples of Data Augmentation and Gaussian Noise application on Screw Bag samples; (a) & (b) Shows samples after Data augmentation; (c) Shows samples after Gaussian Noise.**

the range of [0.9, 1.5], with additional minor zooming in the range of [0.1, 0.2] to preserve critical information. Furthermore, horizontal flipping was applied with a 50% probability. The augmented images were assigned the same class labels as the original images, and the training dataset was expanded, improving the model's generalization ability and accuracy in detecting objects [40]. During training, the original and augmented images were subjected to further in situ data augmentation, where comparable random changes were applied to each batch in every epoch. The continual increase in data enabled the network to experience a wide range of variables, facilitating the creation of a more complex model and preventing the model from overfitting to the training data.

F. GAUSSIAN NOISE-ENHANCED GENERALIZATION

This methodology incorporates the addition of Gaussian noise to both raw and transformed training images to enhance model generalization to new images. This deliberate distortion aims to improve the models' generalization capabilities. Moreno-Barea [41] demonstrated the effectiveness of adding Gaussian noise by testing nine image datasets from the UCI repository and observing improvements in all cases. Introducing Gaussian noise during training enables the model to become more adept at handling new, previously unseen disturbances, thereby improving accuracy. Gaussian noise, which has a Probability Density Function (PDF) similar to the standard Gaussian distribution, is a reliable and valuable method for adding variation to training data. Given is the PDF $f(x)$ of a Gaussian random variable x:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{15}$$

The magnitude of the noise introduced is directly proportional to the standard deviation (σ). The proposed method used a Gaussian distribution with a mean of $0\,(\mu = 0)$ and a standard deviation of 0.05 ($\sigma = 0.05$) to generate noise samples for each image shown in Fig. 6. The magnitude of the noise is kept low to avoid losing significant detail. This value was determined after exploratory analysis to identify the best fit for the data.

G. ATTENTION MECHANISMS FOR FEATURE REFINEMENT

The proposed model employs and integrates three complementary attention modules: CFA, RHA, and GCA, to refine backbone feature maps and enhance both anomaly detection and localization. The CFA produces channel-wise importance weights to emphasize discriminative feature channels. The

RHA generates spatial gating to highlight salient regions while suppressing background noise. The GCA captures long-range dependencies through multi-head self-attention, modeling global interactions between spatial locations and channels.

These attention modules are incorporated into the fine-tuning stage of each backbone. Leveraging transfer learning, early layers are typically frozen to retain general-purpose visual representations [42], while later layers and the inserted attention blocks are trained on the target datasets to adapt feature representations to the anomaly detection task. The refinement begins with attention, followed by two convolutional blocks, each consisting of a convolution operation Conv2D, ReLU activation $\sigma(x) = \max(0, x)$, and batch normalization BatchNorm. The final output is generated by concatenating the outputs of the regression head $f_r(x)$ and the classification head $f_c(x)$, defined as:

$$\text{Output} = \text{Concat}(f_r(x), f_c(x)) \tag{16}$$

where $f_r(x)$ performs anomaly regression and $f_c(x)$ performs classification as illustrated in Fig. 5. Three distinct attention mechanisms are integrated into the model at strategic stages to enhance feature extraction and representation, addressing different aspects of the input features. These mechanisms work together to improve anomaly detection and localization, focusing on channels, spatial regions, and global dependencies. Fig. 5 (b) The CFA emphasizes the most critical feature channels by generating channel-wise attention maps. For an input feature tensor $F \in \mathbb{R}^{H\times W\times C}$, global statistics are computed across spatial dimensions for each channel:

$$F_{\min} = \text{reduce\_min}(F, \text{axis} = (H, W)), \tag{17}$$
$$F_{\max} = \text{reduce\_max}(F, \text{axis} = (H, W)) \tag{18}$$

These statistics are combined using an addition operation, followed by a sigmoid activation to generate channel-wise attention weights:

$$A_c = \sigma(F_{\min} + F_{\max}), \tag{19}$$

where σ represents the sigmoid function. The refined features are then obtained by applying the attention weights through element-wise multiplication:

$$F_{\text{CF}} = F \odot A_c \tag{20}$$

and subsequently passed through a 7×7 convolutional layer Conv2D$(1, (7, 7))$ with stride $(1, 1)$ for further refinement. Fig. 5 (c) The RHA enhances attention to salient spatial regions while filtering irrelevant backgrounds. For input features F, global max-pooling and global average-pooling are applied to extract region-specific statistics:

$$F_{\max} = \text{GlobalMaxPooling2D}(F), \tag{21}$$
$$F_{\text{avg}} = \text{GlobalAveragePooling2D}(F) \tag{22}$$

TABLE 1. Model fine-tuning hyperparameters.

Parameters	Values
Number of classes	5 & 15
Learning rate	0.0001
Batch size	32
Activation	ReLU, Linear, Sigmoid & Softmax
Loss function	Sparse Categorical Crossentropy
Optimizer	Adam
Local training epoch	120

Each pooled feature is passed through two fully connected layers with ReLU activations, followed by a sigmoid activation to generate attention weights:

$$A_{\max} = \sigma(\text{Dense}(\text{ReLU}(\text{Dense}(F_{\max})))), \tag{23}$$
$$A_{\text{avg}} = \sigma(\text{Dense}(\text{ReLU}(\text{Dense}(F_{\text{avg}})))) \tag{24}$$

These weights are combined and applied to the input features through element-wise multiplication, enhancing focus on salient regions:

$$F_{\text{RH}} = F \odot (A_{\max} + A_{\text{avg}}) \tag{25}$$

Fig. 5 (d) the GCA captures long-range dependencies and complex interactions between spatial locations and feature channels. For an input feature tensor F, we first compute the query, key, and value representations via learned linear projections:

$$Q = W_Q F, \quad K = W_K F, \quad V = W_V F \tag{26}$$

where W_Q, W_K, and W_V are learnable weight matrices that project the input features into the query, key, and value spaces, respectively. The multi-head self-attention output is then given by

$$F_{\text{MHA}} = \text{MHA}(Q, K, V) \tag{27}$$

which corresponds to applying multi-head self-attention to different projections of the same input tensor F. The output of the multi-head attention is added to the original input via a residual connection and normalized using layer normalization:

$$F_{\text{GC}} = \text{LayerNorm}(F_{\text{MHA}} + F) \tag{28}$$

These attention mechanisms collectively improve the model's ability to extract meaningful features for anomaly detection. The CFA highlights critical feature channels, the RHA refines attention to salient spatial regions, and the GCA accounts for global dependencies, ensuring precise and robust performance across diverse anomaly conditions.

H. TRAINING AND HYPERPARAMETER OPTIMIZATION

This study fine-tunes three widely used CNN backbones: VGG16, DenseNet121, and ResNet50, augmented with attention modules, and adopts hyperparameter settings suitable for industrial anomaly-detection scenarios. Each backbone is adapted via targeted fine-tuning, and the chosen parameter values are summarized in Table 1. Hyperparameters (learning rate, batch size, optimizer, loss function, and number of epochs) were established based on small-scale validation experiments and commonly accepted best-practice values in the literature; these settings yielded stable convergence and consistent validation performance on the MVTec datasets. The configuration provides an effective and high-performance framework for defect detection in various industrial environments.

The training process in these experiments is managed using the Keras Callback tool, which enables the execution of procedures at the beginning or end of an epoch or training batch. These callbacks allow monitoring of the model's internal states and data throughout training, making it easier to visualize and export progress. They also store models at regular intervals, stop training early to prevent overfitting, and adjust the learning rate when a metric stops improving. The vital functions utilized are EarlyStopping, which halts training if the validation error fails to decrease after a specified number of epochs; ModelCheckpoint, which saves the best model based on validation errors; and ReduceLROnPlateau, which reduces the learning rate by a factor of 0.5 if the validation error does not improve after 25 epochs. Additionally, TensorBoard was utilized to display the training progress, while CSVLogger was employed to export the training data to a CSV file for future analysis.

We set a maximum of 150 epochs for optimal model convergence, with a minimum learning rate of 0.0001 since lower values did not enhance performance. A Gaussian noise variance of 0.3 was selected based on visual inspection to ensure the images were noisy yet recognizable. The parameters for the optimizer were set as follows: $\alpha = 10^{-4}$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-7}$. While some parameters are dataset-dependent, these settings provide a general methodology applicable to various problems. For the MVTec LOCO AD [36] and MVTec AD [37] datasets, preprocessing steps included resizing the images to 224×224 pixels, which is suitable for transfer learning CNN models. The overall pipeline of model training is illustrated in Fig. 7, which comprises several steps, including normalization, standardization, image generation, dataset splitting, data augmentation, Gaussian noise injection, model fine-tuning, model training, and evaluation.

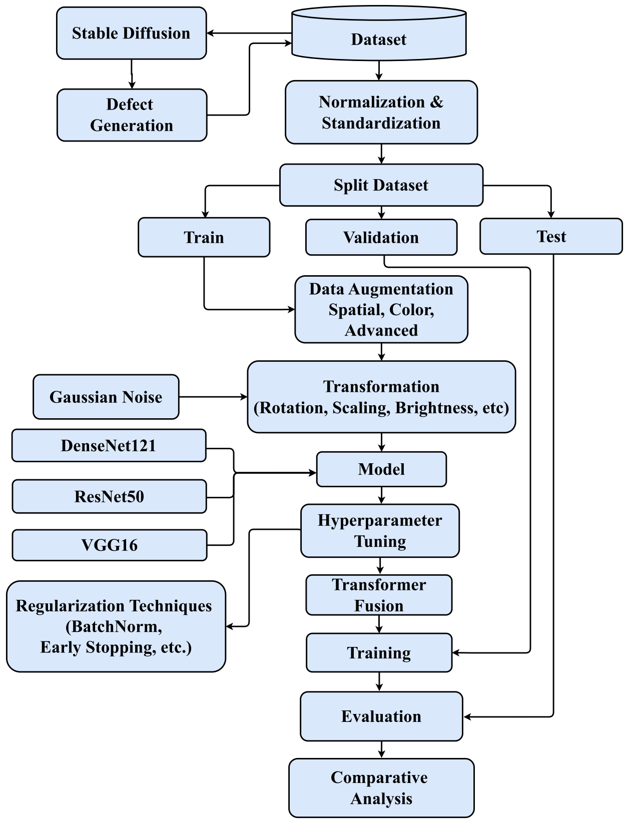

FIGURE 7. Overall model training pipeline.

IV. EXPERIMENTAL SETUP

A. DATASET DESCRIPTION

The dataset contains 3644 images of 5 classes from the MVTec Logical Constraints Anomaly Detection (MVTec LOCO AD) [36] dataset and 5000 images of 15 classes from the MVTec Anomaly Detection (MVTec AD) [37] dataset. Inspired by real-world industrial inspection scenarios, these datasets comprise images from various categories that encompass both structural and logical anomalies. Logical defects such as scratches, dents, color spots, or cracks, and structural defects like misplacement or missing parts, these defective appearances vary in different sizes, shapes, and types, and in most cases, only contain a small fraction of anomalous pixels. This makes it extremely difficult to identify when there is a limited number of anomaly samples. We generated 1,520 anomaly images with a resolution of (256×256) for each category. Examples of some of the categories from both datasets are shown in Fig. 4 and Fig. 9.

B. SIMULATION ENVIRONMENT

The study utilized a machine equipped with an Intel Core i9-10980XE processor at 3 GHz, three NVIDIA GeForce RTX 3090 GPUs, and 130 GB of RAM, running Ubuntu 22.04.4 LTS for optimal performance. The deep learning software setup utilized driver version 535.161.08 for the server, which was chosen for its GPU capability and compatibility with CUDA version 12.2.0. Python versions 3.8 and 3.9 were utilized for the TensorFlow and PyTorch setup.

C. EVALUATION METRICS

This paper employs multiple performance metrics to comprehensively evaluate and compare the proposed method against existing approaches, in addition to conventional detection metrics, including Accuracy, Precision, F1-score, Recall, Specificity, and Area Under the Receiver Operating Curve (AUROC). To calculate the AUROC, operating points were derived from various detection thresholds, and then the TPR

TABLE 2. Performance of CNN models for different classes on MVTec LOCO AD [36].

Class Name	Best Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Breakfast box	DenseNet121	87.37	87.20	86.45	86,82
Pushpins	VGG16	88.10	88.12	87.01	87.56
Juice bottle	ResNet50	88.85	88.11	87.08	87.59
Splicing connectors	ResNet50	86.56	86.34	85.09	85.71
Screw bag	VGG16	87.45	86.05	87.14	86.59

TABLE 3. Performance of CNN models for different classes on MVTec AD [37].

Class Name	Best Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Bottle	VGG16	88.76	87.23	88.36	87.79
Cable	DenseNet121	88.32	88.54	87.23	87.88
Capsule	VGG16	89.57	88.85	89.12	88.98
Grid	ResNet50	89.87	89.85	89.23	89.54
Hazelnut	ResNet50	88.69	88.62	88.35	88.48
Leather	DenseNet121	87.48	87.24	87.23	87.23
Metal nut	VGG16	88.19	88.21	88.23	88.22
Pill	VGG16	89.39	89.31	88.25	88.78
Screw	ResNet50	86.33	86.23	85.45	85.84
Tile	ResNet50	88.24	88.34	88.23	88.28
Toothbrush	DenseNet121	89.86	91.86	91.21	91.53
Transistor	DenseNet121	88.76	87.56	88.89	88.22
Wood	VGG16	86.98	86.69	85.56	86.12
Zipper	ResNet50	87.87	87.25	87.32	87.28

TABLE 4. Evaluating the selected models against other models on MVTec LOCO AD and MVTec AD datasets.

Models	MVTec LOCO AD [36]					MVTec AD [37]				
	Accuracy	Precision	Recall	F1	Specificity	Accuracy	Precision	Recall	F1	Specificity
Xception + DenseNet121	90.45 ±0.18	87.29 ±0.17	89.37 ±0.19	88.35 ±0.16	90.67 ±0.17	90.10 ±0.19	88.98 ±0.18	89.95 ±0.20	89.46 ±0.17	90.46 ±0.19
VGG16 + InceptionResNetV2	89.76 ±0.16	90.37 ±0.15	90.98 ±0.17	90.67 ±0.18	88.34 ±0.16	90.89 ±0.15	88.64 ±0.16	90.56 ±0.17	89.59 ±0.18	89.39 ±0.16
EfficientNetB0 + MobileNet + InceptionV3	91.56 ±0.19	91.78 ±0.19	90.63 ±0.18	91.20 ±0.17	90.61 ±0.19	90.81 ±0.18	91.28 ±0.17	91.87 ±0.19	91.57 ±0.20	90.92 ±0.19
DenseNet121 + EfficientNetB0 + NASNet	91.27 ±0.17	90.88 ±0.18	90.49 ±0.16	90.68 ±0.19	90.47 ±0.18	91.59 ±0.16	91.56 ±0.18	90.82 ±0.17	91.19 ±0.19	91.68 ±0.17
ResNet101 + EfficientNetB0 + MobileNetV2 + VGG19	91.69 ±0.18	90.32 ±0.17	92.56 ±0.19	91.43 ±0.21	91.24 ±0.18	92.89 ±0.19	91.24 ±0.16	91.17 ±0.18	91.20 ±0.20	91.85 ±0.17
VGG16 + DenseNet121 + ResNet50	**92.09 ±0.19**	**92.15 ±0.18**	**92.91 ±0.17**	**92.53 ±0.17**	**92.39 ±0.18**	**92.68 ±0.19**	**92.51 ±0.17**	**92.64 ±0.20**	**92.57 ±0.20**	**92.98 ±0.19**

and FPR were computed at each threshold, where each threshold corresponds to a distinct predicted score. Here, $(\mathrm{FPR}_i, \mathrm{TPR}_i)$ for $i = 0, \ldots, M$ be the pairs of FPR and TPR computed at these thresholds, with $(\mathrm{FPR}_0, \mathrm{TPR}_0) = (0, 0)$ and $(\mathrm{FPR}_M, \mathrm{TPR}_M) = (1, 1)$. The AUROC is approximated using the trapezoidal rule:

$$\mathrm{AUROC} \approx \sum_{i=0}^{M-1} (\mathrm{FPR}_{i+1} - \mathrm{FPR}_i) \frac{\mathrm{TPR}_i + \mathrm{TPR}_{i+1}}{2} \quad (29)$$

The study also reports key computational and resource-oriented measures. These include the number of epochs, Floating-Point Operations per Second (FLOPs), inference time (in milliseconds), time per epoch (in seconds), and overall model size (in megabytes). A more holistic understanding of each model's trade-offs and practical utility is achieved by incorporating both performance and complexity metrics. Furthermore, a detailed class-wise performance breakdown is provided for the best-performing models, supplemented by confusion matrices that offer insights into the nature and frequency of misclassification.

V. EXPERIMENTAL RESULTS AND DISCUSSION

A. CNN MODELS PERFORMANCE ANALYSIS

The CNN models, each integrated with specialized attention mechanisms, were evaluated on the MVTec LOCO AD [36] and MVTec AD [37] datasets to assess their performance on individual defect classes. Tables 2 and 3 present a detailed breakdown of the class-wise results. Notably, there is no

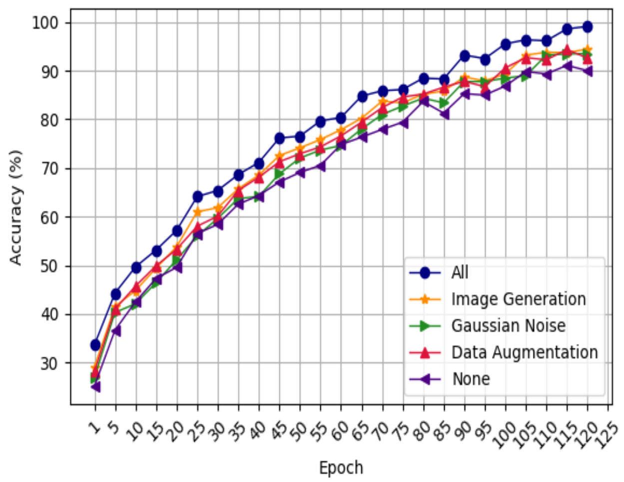

FIGURE 8. **Impact of various image processing methods on MCTNet model accuracy across epochs.**

single model that dominates across all classes. DenseNet121 excels at detecting anomalies in the *Breakfast box* category, while VGG16 is the top choice for *Pushpins*. Similarly, ResNet50 shows superior results for *Juice bottle* and *Splicing connectors*. A comparable trend emerges in the MVTec AD dataset, where different models claim superiority for different classes, VGG16 for *Bottle* and *Capsule*, DenseNet121 for *Cable* and *Leather*, and ResNet50 for *Grid* and *Hazelnut*.

These observations highlight the inherent diversity in defect types and the complexity of developing a single model that can handle all variations equally well. Certain architectures may be more effective at capturing subtle textural anomalies, while others may be more capable of identifying geometric distortions. This class-dependent performance variability thus highlights the necessity for a fusion-based approach, where the complementary strengths of multiple models and attention mechanisms are combined. This observation influenced the proposal of MCTNet, a Multi-CNN Transformer fusion model that integrates the best features from each model, offering a more robust and consistently high-performing framework for industrial defect detection. To further validate the selection of VGG16, DenseNet121, and ResNet50, additional fusion-based evaluations were conducted on several other models, as summarized in Table 4. Models such as MobileNetV2, EfficientNetB0, Xception, InceptionResNetV2, NASNet, and InceptionV3 generally exhibited lower performance, due to architectural constraints that limit their feature extraction capabilities.

B. IMPACT OF IMAGE GENERATION AND PREPROCESSING

The proposed advanced industrial anomaly detection system integrates advanced image processing techniques to overcome the limitations of existing state-of-the-art anomaly detection approaches, particularly those arising from dataset constraints. To address issues such as insufficient data diversity, class imbalance, and the inherent difficulty of capturing rare industrial defects, the framework employs a combination of data augmentation, Gaussian noise injection, and diffusion-based image synthesis. By enriching the training set with a broader and more representative range of defective samples, the model gains robustness and improves its ability to generalize beyond the original data distribution. As illustrated in Fig. 8, the introduced augmentation, Gaussian noise, and image generation lead to a more accelerated and stable increase in model accuracy throughout training. The figure demonstrates that combining multiple image processing strategies and generative image creation techniques yields a synergistic effect, enabling more effective and consistent anomaly detection in complex industrial scenarios.

C. MCTNET PERFORMANCE COMPARISON

The proposed transformer-based fusion model that integrates feature representations from multiple attention-empowered CNN architectures, shows accurate anomaly localization performance in Fig. 9 and demonstrates significant improvements in defect detection performance compared to existing methods, as illustrated in Table 5. This table provides a comparative analysis against existing Fusion approaches, including CRCFusionAICADx [43], which employs VGG16, DenseNet201, and ResNet50 for diverse feature extraction and utilizes CNN + LSTM layers for detection; CSFNet (Cross-scale Feature Fusion) [44], leveraging ResNet34 to enhance feature representation across multiple scales; IEFCM (Iterative Enhancement Fusion-based Cascaded) [45], applying iterative enhancement techniques within a cascaded framework to progressively refine feature fusion; Multi-branch Fusion [46], integrating YOLO with a Feature Directional Selection Module and Multi-Branch Fusion Attention to effectively combine features from multiple detection branches; and Attention-based Fusion [47], utilizing attention mechanisms to dynamically weigh and merge features from various sources.

On the MVTec LOCO AD [36] dataset, MCTNet achieves an accuracy of 98.56%, precision of 98.08%, recall of 97.41%, and an F1-score of 97.74%, surpassing all other methods. Models like CSFNet, and Attention-based Fusion demonstrated an accuracy of 95.12% and 95.79%. MCTNet's superior performance underscores its effective synergy in leveraging features derived from multiple CNN backbones. Similarly, on the MVTec AD [37] dataset, MCTNet attains an accuracy of 97.78%, precision of 96.84%, recall of 96.84%, and an F1-score of 96.84%, outperforming all compared fusion strategies. In contrast, CRCFusionAICADx, CSFNet, IEFCM, Multi-branch Fusion, and Attention-based Fusion achieve competitive yet lower F1-scores of 94.89%, 92.64%, 93.63%, 94.82%, and 95.56%, respectively. Furthermore, MCTNet combines three CNN backbones with a transformer-based meta-learner; its computational cost remains comparable to existing fusion frameworks. Specifically, MCTNet requires 15.3 GFLOPs per image, which is the same as CRCFusionAICADx and within the same order of magnitude

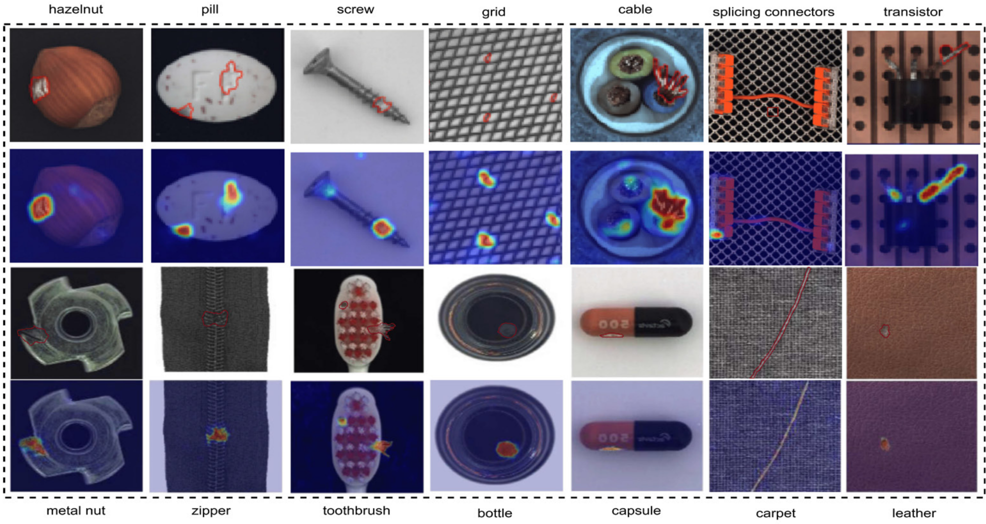

FIGURE 9. Visualization of anomaly detection results across both datasets: The first row displays the anomalous samples, while the second row highlights the model's attention on the detected anomalies.

TABLE 5. Evaluation of the proposed method against existing approaches on MVTec LOCO AD and MVTec AD datasets.

Model	MVTec LOCO AD [36]				MVTec AD [37]				Backbone	Epoch	FLOPs (GFLOPs)	Inf. Time (ms)	Time/ Epoch (s)	Model Size (MB)
	Acc.	Prec.	Rec.	F1	Acc.	Prec.	Rec.	F1						
CRCFusion AICADx [43]	94.58	93.86	94.41	94.13	95.87	95.24	94.54	94.89	VGG16 DenseNet201 ResNet50	103	15.3	50	51	354.90
Cross-scale Feature Fusion [44]	95.12	95.18	94.58	94.88	93.56	93.04	92.24	92.64	ResNet34	89	8.7	45	48	291.78
IEFCM [45]	94.42	93.73	93.53	93.83	94.24	94.20	93.06	93.63	CNN	95	7.2	37	42	212.12
Multi-branch Fusion [46]	96.15	95.08	96.19	95.63	95.89	95.54	94.12	94.82	YOLO	110	9.8	50	49	317.76
Attention-based Fusion [47]	95.79	95.31	95.17	95.24	95.79	95.70	95.42	95.56	ResNet50	98	11.1	48	49	299.98
Multi-CNN Transformer Fusion (Proposed)	**98.56**	**98.08**	**97.41**	**97.74**	**97.78**	**96.84**	**96.84**	**96.84**	**VGG16 DenseNet121 ResNet50**	**120**	**15.3**	**45**	**48**	**319.20**

TABLE 6. Image-level AUROC performance comparison with existing studies.

Model	MVTec LOCO AD [36]	MVTec AD [37]	Average
Patchcore [19]	99.1	79.3	89.2
PaDiM [48]	95.3	68.9	82.1
SPADE [49]	85.5	68.8	77.2
GCAD [36]	93.1	83.3	88.2
S-T [50]	93.2	77.4	85.3
TUT [51]	98.9	90.1	94.5
MCTNet (Proposed)	**98.15**	**97.56**	**97.85**

as CSFNet (8.7 GFLOPs) and IEFCM (7.2 GFLOPs), while delivering notably higher F1-scores on both MVTec LOCO AD and MVTec AD. The inference time of 45 ms is equal to or smaller than that of four out of the five baselines, and the training time per epoch (48 s) is similar to CSFNet and Attention-based Fusion. In terms of memory footprint, the model size of 319.20 MB is smaller than CRCFusionAICADx (354.90 MB) and comparable to Multi-branch Fusion and Attention-based Fusion. This reflects a balanced trade-off between computational complexity and performance, making MCTNet a highly effective solution for defect detection tasks. The classwise detection performance of MCTNet is presented in Fig. 10, where for MVTec AD, classes like capsule, wood, and zipper achieve over 95% accuracy, with minimal misclassifications distributed across other classes. Similarly, for MVTec LOCO AD, classes such as pushpins and juice_bottle achieve near-perfect detection with 99% and 98% accuracy, respectively. The misclassifications are sparse and limited, further validating MCTNet's efficiency. Combined with its top-tier detection metrics, MCTNet establishes itself as a highly effective and practical solution compared to existing fusion-based methods.

D. COMPARISON WITH EXISTING APPROACHES

Fig. 11 illustrates the AUROC curves for the MCTNet's performance in anomaly detection across five distinct categories of the MVTec LOCO AD [36] dataset. The highest AUC

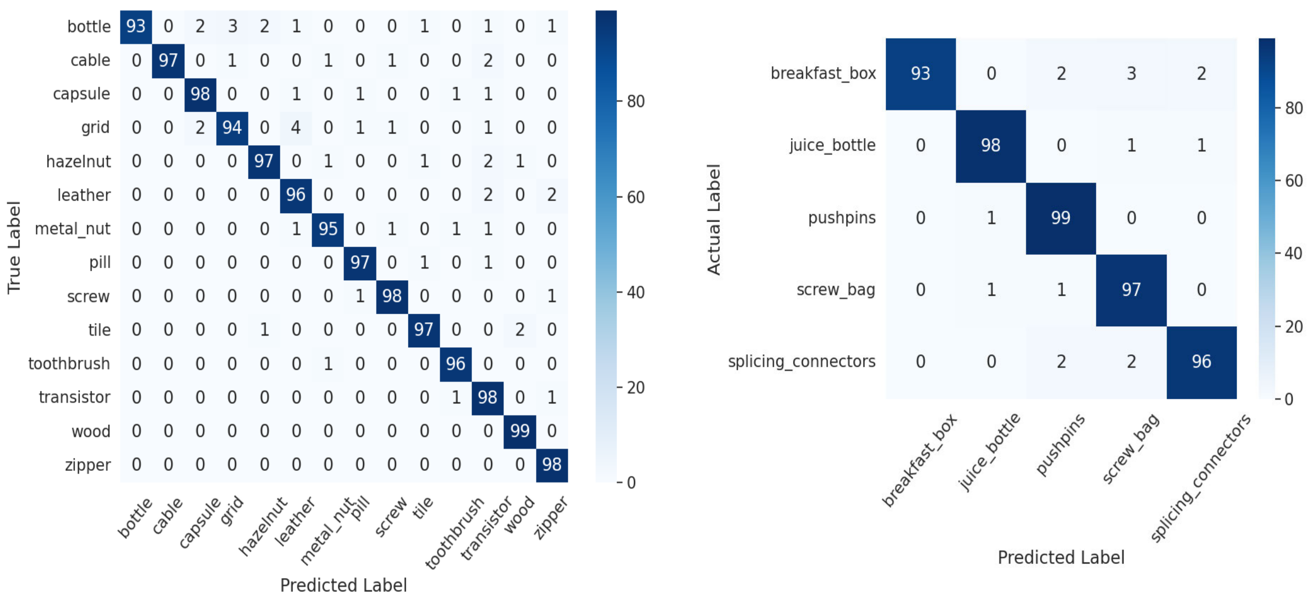

FIGURE 10. Normalized confusion matrix of highest accuracy on each class of MVTec AD [37] and MVTec LOCO AD [36] datasets using MCTNet.

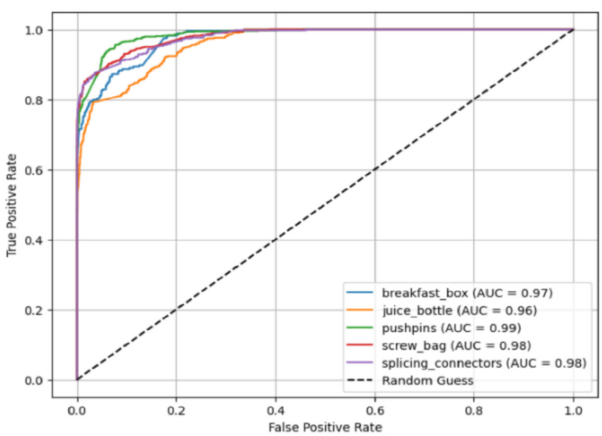

FIGURE 11. AUROC performance on the MVTec LOCO AD [36] dataset.

value is 0.99 for *Pushpins*, and the lowest is 0.96 for both *juice_bottle*. This demonstrates the model's strong discriminative ability. These results highlight MCTNet's effectiveness in distinguishing between normal and anomalous instances, maintaining a high true positive rate while minimizing false positives, and reinforcing its robustness and reliability for industrial anomaly detection tasks. Table 6 presents a performance comparison of AUROC percentages between the proposed model and existing studies on the MVTec LOCO AD and MVTec AD [37] datasets. The table highlights that the proposed model achieves an AUROC of 98.15% on MVTec LOCO AD and 97.56% on MVTec AD, resulting in an average AUROC of 97.85%. This performance surpasses other methods such as Patchcore [19], which shows 99.1% on MVTec LOCO AD but drops to 79.3% on MVTec AD with an average of 89.2%. PaDiM [48], SPADE [49], and GCAD [36] show lower AUROC percentages across both datasets, while TUT [51] performs competitively with an average of 94.5%. The results indicate that the proposed model demonstrates superior and consistent performance across different datasets compared to existing methods.

VI. CONCLUSION

This paper introduces MCTNet, a transformer-based fusion model that integrates feature representations from multiple attention-empowered CNN architectures, specifically VGG16, DenseNet121, and ResNet50. By addressing challenges such as class imbalance, limited data variability, and model overfitting through the VARSDiffusion model for synthetic image generation, extensive data augmentation, and Gaussian noise injection, MCTNet effectively improves the generalization and robustness of anomaly detection in manufacturing environments. Empirical evaluations on the MVTec LOCO AD [36] and MVTec AD [37] datasets demonstrate that MCTNet achieves superior performance, with AUROC scores of 98.15% and 97.56%, respectively. These results surpass existing state-of-the-art fusion methods, highlighting the effectiveness of combining multiple CNN backbones with a Transformer-based meta-learner for feature fusion. Additionally, MCTNet maintains computational efficiency with a balanced trade-off between FLOPs, inference time, and model size. The proposed approach not only enhances detection accuracy and reliability but also ensures robustness across diverse anomaly types. Despite the strong anomaly detection performance, the proposed MCTNet has some limitations. The architecture combines three high-capacity CNN backbones with a transformer-based meta-learner, leading to a comparatively large number of parameters and increased computational and memory demands. This makes the full model less suitable for deployment on extremely resource-constrained edge devices. In addition, the current evaluation is restricted to publicly available industrial anomaly detection benchmarks and controlled experimental conditions; the method has not yet been deployed or validated on real industrial inspection systems or embedded edge hardware. Future work will therefore focus on model compression and hardware-aware optimization, as well as comprehensive on-device evaluation in real-world industrial environments.

ACKNOWLEDGMENT

The authors would like to thank King Fahd University of Petroleum and Minerals, Saudi Arabia, for their research collaboration.

REFERENCES

[1] W. Luo, H. Yao, and W. Yu, "Normal reference attention and defective feature perception network for surface defect detection," *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–14, 2023.

[2] J. Shin, M. Hamza, H. Kim, J. Jo, B. Park, and Y. Kim, "Lightweight Swin transformer for high-precision industrial defect detection in smart manufacturing," *Alexandria Eng. J.*, vol. 130, pp. 227–240, Oct. 2025.

[3] J. H. Cavalcanti, T. Kovacs, A. Ko, and K. Pocsarovszky, "Production system efficiency optimization through application of a hybrid artificial intelligence solution," *Int. J. Comput. Integr. Manuf.*, vol. 37, no. 6, pp. 790–807, Jun. 2024.

[4] H. Dong, K. Song, Y. He, J. Xu, Y. Yan, and Q. Meng, "PGA-Net: Pyramid feature fusion and global context attention network for automated surface defect detection," *IEEE Trans. Ind. Informat.*, vol. 16, no. 12, pp. 7448–7458, Dec. 2020.

[5] K. Vinod, A. Sasinthiran, N. Thangaraj, and S. Gnanasekaran, "A hybrid deep learning approach to automate industrial quality control," *Int. J. Comput. Integr. Manuf.*, vol. 2025, pp. 1–21, Jul. 2025.

[6] M. Abbas and M. Shafiee, "Structural health monitoring (SHM) and determination of surface defects in large metallic structures using ultrasonic guided waves," *Sensors*, vol. 18, no. 11, p. 3958, Nov. 2018.

[7] P. K. Mishra, A. Mihailidis, and S. S. Khan, "Skeletal video anomaly detection using deep learning: Survey, challenges, and future directions," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 8, no. 2, pp. 1073–1085, Apr. 2024.

[8] Y. Cui, Z. Liu, and S. Lian, "A survey on unsupervised anomaly detection algorithms for industrial images," *IEEE Access*, vol. 11, pp. 55297–55315, 2023.

[9] B. Staar, M. Lütjen, and M. Freitag, "Anomaly detection with convolutional neural networks for industrial surface inspection," *Proc. CIRP*, vol. 79, pp. 484–489, Jan. 2019.

[10] S. Samal, Y.-D. Zhang, J. M. G. Saez, S.-H. Wang, B. K. Balabantaray, and R. Nayak, "EAYv3-CFC^3: Ensemble learning with attention-based Yv3 combined with CFC^3 loss for obscenity detection," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 8, no. 2, pp. 1097–1101, Apr. 2024.

[11] Y. Zhou, C. Guo, X. Wang, Y. Chang, and Y. Wu, "A survey on data augmentation in large model era," 2024, *arXiv:2401.15422*.

[12] N. K. Hosseini, H. Toshani, S. A. Jalebi, and S. Sharifzadeh, "Enhanced bearing fault detection in induction motors using projection-based SVM," *IEEE Trans. Ind. Appl.*, vol. 61, no. 3, pp. 3623–3636, May 2025.

[13] M. Goldstein and S. Uchida, "A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data," *PLoS ONE*, vol. 11, no. 4, Apr. 2016, Art. no. e0152173.

[14] S. M. Acosta, R. M. A. Oliveira, and Â. M. O. Sant'Anna, "Machine learning algorithms applied to intelligent tyre manufacturing," *Int. J. Comput. Integr. Manuf.*, vol. 37, no. 4, pp. 497–507, Apr. 2024.

[15] O. Torkul, I. H. Selvi, and M. Şişci, "Smart seru production system for industry 4.0: A conceptual model based on deep learning for real-time monitoring and controlling," *Int. J. Comput. Integr. Manuf.*, vol. 37, no. 4, pp. 385–407, Apr. 2024.

[16] R.-C. Wu, S. Deng, R. Ran, H.-X. Dou, and L.-J. Deng, "INF^3: Implicit neural feature fusion function for multispectral and hyperspectral image fusion," *IEEE Trans. Comput. Imag.*, vol. 10, pp. 1547–1558, 2024.

[17] N. Cohen and Y. Hoshen, "Sub-image anomaly detection with deep pyramid correspondences," 2020, *arXiv:2005.02357*.

[18] X. Tao, D. Zhang, W. Ma, Z. Hou, Z. Lu, and C. Adak, "Unsupervised anomaly detection for surface defects with dual-siamese network," *IEEE Trans. Ind. Informat.*, vol. 18, no. 11, pp. 7707–7717, Nov. 2022.

[19] K. Roth, L. Pemula, J. Zepeda, B. Schölkopf, T. Brox, and P. Gehler, "Towards total recall in industrial anomaly detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 14298–14308.

[20] P. Xing and Z. Li, "Visual anomaly detection via partition memory bank module and error estimation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 8, pp. 3596–3607, Aug. 2023.

[21] M. Rudolph, B. Wandt, and B. Rosenhahn, "Same same but DifferNet: Semi-supervised defect detection with normalizing flows," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 1906–1915.

[22] B. Zhang, B. Yang, X. Wang, and C. Vian, "Patchflow: Leveraging a flow-based model with patch features," in *Proc. ACCV*, vol. 13847, 2024, pp. 535–551.

[23] Y. Zhou, X. Xu, J. Song, F. Shen, and H. T. Shen, "MSFlow: Multiscale flow-based framework for unsupervised anomaly detection," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 36, no. 2, pp. 2437–2450, Feb. 2025.

[24] H.-B. Zhang, C.-Y. Zhang, D.-J. Cheng, K.-L. Zhou, and Z.-Y. Sun, "Detection transformer with multi-scale fusion attention mechanism for aero-engine turbine blade cast defect detection considering comprehensive features," *Sensors*, vol. 24, no. 5, p. 1663, Mar. 2024.

[25] X. Huang, Y. Li, Y. Bao, and X. Zhu, "Sparse cross-transformer network for surface defect detection," *Sci. Rep.*, vol. 14, no. 1, p. 24731, Oct. 2024.

[26] K. Li, L. Wang, X. Gao, and L. Zhang, "Transformer-enhanced meta-learning for few-shot fault diagnosis of electric submersible pump," *Expert Syst. Appl.*, vol. 284, Jul. 2025, Art. no. 127851.

[27] Z. Zuo, J. Dong, Y. Qu, and Z. Wu, "Training-free anomaly generation via dual-attention enhancement in diffusion model," 2025, *arXiv:2508.11550*.

[28] H. Sun, Y. Cao, H. Dong, and O. Fink, "Unseen visual anomaly generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2025, pp. 25508–25517.

[29] Z. Dai, S. Zeng, H. Liu, X. Li, F. Xue, and Y. Zhou, "SeaS: Few-shot industrial anomaly image generation with separation and sharing fine-tuning," 2024, *arXiv:2410.14987*.

[30] X. Y. Lee, L. Vidyaratne, M. Alam, A. Farahat, D. Ghosh, T. G. Diaz, and C. Gupta, "XDNet: A few-shot meta-learning approach for cross-domain visual inspection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2023, pp. 4375–4384.

[31] Z. Zhao, H. Guo, and Y. Wang, "A multi-information fusion anomaly detection model based on convolutional neural networks and AutoEncoder," *Sci. Rep.*, vol. 14, no. 1, p. 16147, Jul. 2024.

[32] Y. Liu, L. Zhu, L. Ding, Z. Huang, H. Sui, S. Wang, and Y. Song, "Selective ensemble method for anomaly detection based on parallel learning," *Sci. Rep.*, vol. 14, no. 1, p. 1420, Jan. 2024.

[33] P. Banerjee, V. Sivaramakrishnan, and A. K. Srivastava, "Decentralized modular nonlinear physics-informed neural network (mnPINN) for synchrophasor data anomaly detection," *IEEE Trans. Ind. Appl.*, vol. 61, no. 2, pp. 2490–2503, Mar. 2025.

[34] P. K. Mohanty, P. Jena, and N. P. Padhy, "TimeGAN-based diversified synthetic data generation following BERT-based model for EV battery SOC prediction: A state-of-the-art approach," *IEEE Trans. Ind. Appl.*, vol. 61, no. 3, pp. 4167–4185, Mar. 2025.

[35] Y. Cui, R. Butler, Z. Liu, T. Jaremko, and S. Hinds, "The future of predictive maintenance for variable frequency drives based on AI/ML," *IEEE Trans. Ind. Appl.*, vol. 61, no. 4, pp. 1–9, Jul. 2025.

[36] P. Bergmann, K. Batzner, M. Fauser, D. Sattlegger, and C. Steger, "Beyond dents and scratches: Logical constraints in unsupervised anomaly detection and localization," *Int. J. Comput. Vis.*, vol. 130, no. 4, pp. 947–969, Apr. 2022.

[37] P. Bergmann, M. Fauser, D. Sattlegger, and C. Steger, "MVTec AD—A comprehensive real-world dataset for unsupervised anomaly detection," in *Proc. IEEE/CVF Conf. Comput. Vis. pattern Recognit.*, Jun. 2019, pp. 9592–9600.

[38] Z. Ding, Q. Lu, X. Wei, Z. Xu, X. Li, S. Guo, C. Gong, Y. Hu, and Y. Han, "Forecasting flexible regulation capabilities of electric vehicles: A probabilistic approach with Bayesian neural networks and transformer," *IEEE Trans. Ind. Appl.*, vol. 61, no. 4, pp. 5468–5478, Jul. 2025.
[39] M. M. Alam, M. A. Dini, D.-S. Kim, and T. Jun, "TMNet: Transformer-fused multimodal framework for emotion recognition via EEG and speech," *ICT Exp.*, vol. 11, no. 4, pp. 657–665, Aug. 2025.
[40] A. Dandekar, R. Zen, and S. Bressan, "A comparative study of synthetic dataset generation techniques," in *Proc. 29th Int. Conf. Database Expert Syst. Appl.*, Regensburg, Germany. Cham, Switzerland: Springer, 2018, pp. 387–395.
[41] F. J. Moreno-Barea, F. Strazzera, J. M. Jerez, D. Urda, and L. Franco, "Forward noise adjustment scheme for data augmentation," in *Proc. IEEE Symp. Ser. Comput. Intell. (SSCI)*, Nov. 2018, pp. 728–734.
[42] Z. Liu, J. Yu, B. Yan, and G. Wang, "A deep 1-D CNN and bidirectional LSTM ensemble model with arbitration mechanism for LDDoS attack detection," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 6, no. 6, pp. 1396–1410, Dec. 2022.
[43] A. S. N. Raju, K. Jayavel, T. Rajalakshmi, and M. Rajababu, "CRCFusionAICADx: Integrative CNN-LSTM approach for accurate colorectal cancer diagnosis in colonoscopy images," *Cognit. Comput.*, vol. 17, no. 1, pp. 1–37, Feb. 2025.
[44] L. Wang, C. Long, X. Li, X. Tang, Z. Bai, and H. Gao, "CSFFNet: Lightweight cross-scale feature fusion network for salient object detection in remote sensing images," *IET Image Process.*, vol. 18, no. 3, pp. 602–614, Feb. 2024.
[45] S. Vats, V. Sharma, K. Singh, D. P. Singh, M. Y. Bajuri, D. Taniar, N. Innab, A. Mouldi, and A. Ahmadian, "Iterative enhancement fusion-based cascaded model for detection and localization of multiple disease from CXR-images," *Expert Syst. Appl.*, vol. 255, Dec. 2024, Art. no. 124464.
[46] Y. Zhang, W. Ma, J. Jia, and R. Zhang, "Multi-branch fusion automatic malaria parasites detection algorithm based on YOLO network," *Available SSRN*, Jul. 2024.
[47] Y. Li, S. Zhou, and H. Chen, "Attention-based fusion factor in FPN for object detection," *Int. J. Speech Technol.*, vol. 52, no. 13, pp. 15547–15556, Oct. 2022.
[48] T. Defard, A. Setkov, A. Loesch, and R. Audigier, "PaDiM: A patch distribution modeling framework for anomaly detection and localization," in *Proc. Int. Conf. Pattern Recognit.*, 2021, pp. 475–489.
[49] T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu, "Semantic image synthesis with spatially-adaptive normalization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2332–2341.
[50] P. Bergmann, M. Fauser, D. Sattlegger, and C. Steger, "Uninformed students: Student–teacher anomaly detection with discriminative latent embeddings," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 4182–4191.
[51] Z. Chen, C. Bai, Y. Zhu, and X. Lu, "TUT: Template-augmented U-Net transformer for unsupervised anomaly detection," *IEEE Signal Process. Lett.*, vol. 31, pp. 780–784, 2024.

MD MAHINUR ALAM received the B.S. degree in computer science and engineering from the International University of Business Agriculture and Technology (IUBAT), Bangladesh, in December 2022, and the M.S. degree in IT convergence engineering from the Kumoh National Institute of Technology, South Korea. He is currently pursuing the Ph.D. degree in engineering with Georgia Southern University. Prior to his master's, he was a Software Engineer with Divine IT Ltd., Bangladesh. His research interests include computer vision, federated learning, vision language models (VLMs), and large language models (LLMs).

ABDULAZIZ YAGOUB BARNAWI (Member, IEEE) is currently an Assistant Professor with the Department of Computer Engineering, King Fahd University of Petroleum and Minerals (KFUPM). He is a Research Scholar with the Interdisciplinary Research Center for Intelligent Secure Systems (IRC-ISS), KFUPM. Over the years, he has led several funded research projects as principal investigator while contributing as a team member on others. He has authored and co-authored several peer-reviewed journal and conference publications. His research interests include the Internet of Things (IoT), wireless sensor networks (WSNs), network management, cybersecurity, the Internet of Drones (IoD), and the application of AI/ML in networking.

MOHTASIN GOLAM received the M.Sc. degree in IT convergence engineering from the Kumoh National Institute of Technology, Republic of Korea, in 2021, and the Ph.D. degree from the Department of IT Convergence Engineering, Kumoh National Institute of Technology, in 2025. He is currently a Postdoctoral Research Fellow with the ICT-Convergence Research Center, Kumoh National Institute of Technology. His research interests include deep learning, large language models (LLMs), UAV networks, the IoT, and blockchain.

KANITA JERIN TANHA received the B.S. degree in computer science and engineering from the International University of Business Agriculture and Technology (IUBAT), Bangladesh, in December 2022. She is currently pursuing the M.S. degree in IT convergence engineering with the Kumoh National Institute of Technology, South Korea. Prior to her master's, she was a Software Quality Assurance Engineer with the Right Brains Solutions Ltd., Bangladesh. Her research interests include machine learning, deep learning, digital twin, and intrusion detection.

TAESOO JUN (Member, IEEE) received the B.S. degree in electrical engineering and the M.S. degree from Seoul National University, in February 1998 and 2000, respectively, and the Ph.D. degree in computer engineering from The University of Texas at Austin, in December 2009. Demonstrating his commitment to research and development, he was the Director and a Principal Engineer for the SW Platform Team/Global AI Center with Samsung Research, Samsung Electronics, Seoul, South Korea, from 2010 to 2022. Since March 2022, he has been contributing to academia as an Assistant Professor with the Department of Computer Software Engineering, Kumoh National Institute of Technology. His diverse research interests encompass distributed computing in smart environments, intelligent systems for pervasive computing, AI applications, SW platforms for intelligent systems, and real-time systems.