Pattern Recognition Letters 194 (2025) 32–40

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Enhancing random surface anomaly detection in real-world using a four-stage one-class approach

Pulin Li [a], Guocheng Wu [a], Yanjie Zhou [a,*], Jiewu Leng [b]

[a] School of Management, ZhengZhou University, Zhengzhou, Henan, 450001, China
[b] State Key Laboratory of Precision Electronic Manufacturing Technology and Equipment, Guangdong University of Technology, Guangzhou, 510006, China

ARTICLE INFO

Edited by Maria De Marsico.

Keywords:
Industrial anomaly detection
Unsupervised learning
Adapter tuning
Mechanical manufacturing
Real-world applications

ABSTRACT

Defect detection and localization are critical for quality control in manufacturing, yet existing algorithms and models trained on laboratory datasets often fail in real industrial scenarios due to their static nature, especially in non-mass production. Moreover, limited and heterogeneous defective samples, coupled with costly human annotation, highlight the need for unsupervised methods relying solely on normal images. To address these challenges, we propose the Random Surface Anomaly Detection (RSAD) model, a four-stage one-class anomaly detection and localization approach. Initially, leveraging embedding-based techniques, we introduce transfer learning with a pretrained ImageNet network in extracting locally aggregated features. Next, adapter tuning is applied to transfer these features into the industrial domain, reducing bias towards natural images. Additionally, random Gaussian noise is introduced into normal feature representations within the feature space and a discriminator then scores feature normality. Finally, experiments on the MPDD dataset and other benchmarks, demonstrate the RSAD model's state-of-the-art (SOTA) performance in anomaly detection, validating its trustworthiness in real-world manufacturing environments.

1. Introduction

In the context of real-world mechanical manufacturing and production, maintaining strict product conformity is essential in the workshop. Traditionally, defect detection has relied on manual inspection, which is both inefficient and costly. The subjectivity of human judgment introduces variability that can undermine the entire quality assurance process. Furthermore, differences in prior knowledge and skill levels, along with fatigue from repetitive tasks, reduce the reliability of human-based quality control [1]. To overcome the limitations of manual inspection, deep learning-driven visual intelligence technology has emerged as a transformative solution in quality control. This approach uses advanced deep learning techniques to analyze images and data, enabling automatic defect detection and ensuring a high level of product consistency.

However, the acquisition of labeled defect samples remains challenging in industrial scenarios, making unsupervised deep learning methods particularly appealing. Unlike supervised learning [2], unsupervised methods rely solely on normal samples for training, which are readily available in modern high-optimized production lines. These methods follow the principle of one-class novelty detection, where detectors trained on a particular known class are tasked with identifying whether a query example belongs to this class [3]. In industrial scenarios, qualified products are designated as the known class, whereas defective products in query data are expected to be classified as unknown. Thus, vision intelligence driven by unsupervised learning algorithms can autonomously detect a broad range of anomalies without requiring explicit defect labels, further enhancing its reliability.

Among unsupervised approaches, embedding-based methods have demonstrated strong performance in industrial anomaly detection. These methods primarily rely on pretrained networks to extract representative features from images. After feature extraction, various techniques, such as memory banks, one-class classification, normalizing flows, and knowledge distillation, are employed to enhance anomaly discrimination. However, embedding-based methods often suffer from domain bias, as pretrained feature extractors are optimized for natural images rather than industrial environments, limiting their generalization performance in real-world defect detection.

To address this issue, fine-tuning strategies have been explored to adapt pretrained feature extractors to industrial visual domains. Parameter-efficient fine-tuning (PEFT) updates only a subset of parameters while keeping most of the pretrained model frozen, achieving

* Corresponding author.
E-mail addresses: lipulin@zzu.edu.cn (P. Li), wu33learn@163.com (G. Wu), ieyjzhou@zzu.edu.cn (Y. Zhou), jwleng@gdut.edu.cn (J. Leng).

https://doi.org/10.1016/j.patrec.2025.05.002
Received 12 October 2024; Received in revised form 29 March 2025; Accepted 3 May 2025
Available online 4 May 2025

remarkable performance in multiple representative visual tasks. Specifically, advanced adapter tuning introduces lightweight adapter modules for task-specific optimization. Adapters adjust feature representations without altering the core parameters, enabling effective domain adaptation with minimal computational overhead. In this work, we adopt Multi-cognitive Visual Adapter (Mona) tuning [4], which freezes the pretrained backbone and trains only small adapter layers inserted at strategic positions. This approach preserves pretrained general knowledge while enabling task-specific customization, making it particularly suitable for industrial anomaly detection.

While deep learning has shown promise across various fields, its success relies heavily on the quality and diversity of training data [5]. Many existing industrial defect detection datasets are collected in controlled environments, lacking the variability of practical production scenarios. These simplified lab datasets lead to models that struggle with real-world complexities like irregular surfaces, varying backgrounds, diverse lighting conditions, and motion blur from moving parts. To tackle this, our study focuses on the MPDD (Metal Parts Defect Detection) dataset [6], which accurately reflects the challenges faced in human-operated production lines. To further validate the model's generalizability, additional evaluations are conducted on three representative industrial anomaly detection benchmarks.

Thus, we present a novel model, Random Surface Anomaly Detection (RSAD), tailored for detecting workpiece defects in industrial images captured under real-world conditions. As a one-class classification framework, RSAD integrates four stages: Patch Feature Extractor, Feature Aligner, Non-deterministic Defect Feature Fuser, and Feature Discriminator. Notably, the Feature Aligner adopts an efficient adapter tuning strategy to facilitate domain adaptation while preserving the pretrained feature extractor. Experimental results demonstrate that RSAD achieves SOTA anomaly localization performance with a 98.7 % P-AUROC on the MPDD dataset as well as the new SOTA for 4 out of 6 classes on anomaly detection. Furthermore, Competitive performance across multiple industrial anomaly detection benchmarks shows strong generalization ability of RSAD.

2. Literature review

This section reviews unsupervised learning methods commonly used in industrial defect detection. Additionally, it discusses adapter tuning, a parameter-efficient fine-tuning strategy that preserves pretrained knowledge while enhancing task-specific adaptation, achieving SOTA performance across multiple visual tasks.

2.1. Unsupervised learning in industrial defect detection

Unsupervised methods model the nominal distribution and detect deviations as anomalies, making unsupervised defect detection essentially an anomaly detection problem. These approaches can be mainly categorized into three types, i.e., the reconstruction-based methods, the synthesizing-based methods, and the embedding-based methods.

Reconstruction-based methods aim to reconstruct normal, anomaly-free data and use the reconstruction error to detect anomalies. They assume that anomalous regions cannot be accurately restored from the learned normal patterns. Early approaches typically employed Autoencoders [7] or Generative Adversarial Networks (GANs) [8,9] for image reconstruction. However, these models often face the identical shortcut problem, where abnormal inputs are also well reconstructed, limiting their effectiveness.

To overcome these limitations, recent methods combine advanced architectures, such as Transformer-based models and Mamba, with specialized mechanisms to strengthen representation learning and enhance reconstruction adaptability. For instance, MambaAD pioneers the application of Mamba to multi-class anomaly detection [10]. Building on stronger representations, DMAD introduces a dual memory bank to separately store normal and abnormal patterns, explicitly enhancing feature discrimination under multi-class conditions [11]. Focusing further on reconstruction adaptability, DDAD employs conditioned denoising diffusion guided by target inputs to produce anomaly-free reconstructions [12], while GLAD integrates global-local adaptive diffusion to dynamically refine reconstruction and better handle complex anomalies [13].

This trend toward adaptive reconstruction naturally extends to the more challenging setting of multi-class anomaly detection (MCAD) [10]. MCAD, introduced by Uniad, requires models to identify anomaly across multiple object categories [14]. This introduces new challenges, including class-aware feature learning, inter-category variability, and the risk of identical shortcuts. Recent works have focused on developing reconstruction-based approaches tailored for MCAD. Representative methods, such as DiAD [15], ViTAD [16] and InvAD [17], incorporate advanced mechanisms like diffusion models, transformer-based architectures, and GAN inversion to improve anomaly detection across multiple categories. These efforts highlight the growing trend toward adaptive and category-aware reconstruction models for MCAD.

Synthesizing-based methods generate pseudo-anomalies by perturbing normal samples and train models to restore them to anomaly-free versions. Early approaches typically relied on simple data augmentation strategies, such as patch rearrangement or basic perturbations [18], to synthesize anomalies, but these methods often lack realism and diversity [19]. To overcome these limitations, more recent methods incorporate advanced synthesis mechanisms and discriminative designs. A representative example, DRAEM, integrates reconstruction and discriminative learning, enabling direct anomaly localization without post-processing [20]. Building on this, RealNet addresses the challenge of synthesizing realistic and diverse anomalies by introducing strength-controllable diffusion-based anomaly generation, coupled with adaptive feature and residual selection [21]. Further advancing reconstruction quality, DiffusionAD reformulates the process as a noise-to-norm diffusion, applying Gaussian perturbations to anomalous regions and restoring them through a fast, single-step denoising [22].

Embedding-based methods have recently achieved SOTA performance in image anomaly detection and localization tasks. Embedding-based approaches leverage pretrained models to extract representative features and identify anomalies by analyzing their distribution in the embedding space. Typical approaches include feature distribution modeling [23], memory-bank techniques, and teacher-student frameworks. Early approaches model normal features using multivariate Gaussian assumptions or encoder-decoder structures, but often struggle to capture fine-grained patterns and contextual dependencies [24]. Recent efforts focus on enhancing feature expressiveness and addressing these limitations. For instance, PatchCore focuses on retaining as much local patch-level context as possible by constructing a compact yet representative memory bank, balancing detection performance and inference efficiency [25]. Recognizing the tendency of student models to forget typical normal patterns, MemKD introduces a normality recall memory within a knowledge distillation framework, explicitly guiding feature learning to mitigate this forgetting issue [26]. Complementing these memory-based approaches, CFLOW-AD departs from memory structures entirely, instead modeling feature likelihoods via conditional normalizing flows, which enables real-time anomaly detection with significantly reduced computational overhead [27].

2.2. Adapter tuning in visual model adaptation

Adapter tuning introduces trainable lightweight adapter modules while keeping pretrained backbone frozen, which saves plenty of training costs. This approach preserves pretrained general knowledge while allowing task-specific customization, achieving great success across multiple representative visual tasks [28]. In action recognition, AdaptFormer integrates a lightweight adaptation module, improving the transferability of Vision Transformers with only 1.5 % additional parameters, while still surpassing full fine-tuning [29]. In image

classification, KAdaptation optimizes subspace training through Kronecker [30]. decomposition, achieving a superior accuracy-efficiency trade-off [30]. For dense prediction tasks such as object detection and segmentation, LoRand applies low-rank adaptation, tuning only 1 %−3 % of backbone parameters while maintaining comparable performance to full fine-tuning [31]. Extending adapter tuning further, Mona introduces multi-scale visual filters and normalization layers, making it the first method to outperform full fine-tuning in instance and semantic segmentation, achieving a 1 % performance gain on COCO [4].

3. Framework of RSAD model

When focusing solely on anomaly detection, RSAD functions as a specialized one-class classification model, distinguishing specific class objects within a broader dataset by training exclusively on samples from that particular class. As illustrated in Fig. 1(a), RSAD operates through four distinct stages: Patch Feature Extractor, Feature Aligner, Non-deterministic Defect Feature Fuser and Feature Discriminator, each contributing to the model's overall efficacy.

Furthermore, the pseudo-code detailing both the training and testing procedures is shown in Algorithm 1. Specifically, the training process involves these four modules mentioned above. The testing process utilizes Patch Feature Extractor, Feature Aligner, and Feature Discriminator, while omitting the Defect Feature Fuser.

During the training phase, pretrained Patch Feature Extractor extracts patch features from normal images. Subsequently, Feature Aligner, a parameter-efficient fine-tuning strategy, is trained to adjust pretrained features for domain adaptation. This approach follows adapter tuning, freezing the pretrained backbone and training only lightweight adapter layers inserted at strategic positions. The Feature Aligner refines feature representations without altering core parameters, enabling efficient domain adaptation. Unlike directly synthesizing anomalies onto the images, the Non-deterministic Defect Feature Fuser takes a unique approach by fusing random Gaussian noise into the normal feature space. This process generates stochastic defect features that effectively simulate various types of defects found in industrial settings. Finally, a simple Feature Discriminator, constructed with a few layers of Multi-Layer Perceptron (MLP), evaluates whether a given sample meets the qualification criteria, thereby facilitating accurate anomaly detection.

During the testing phase, two critical aspects warrant attention. Firstly, in contrast to the training phase where only normal data is employed, the test dataset comprises diverse defect types, covering a broad spectrum of scenarios encountered in real-world industrial applications. Secondly, it is essential to highlight the exclusion of the third stage, the Non-deterministic Defect Feature Fuser. By removing this stage, the remaining modules can be seamlessly integrated into an end-to-end network, thereby enabling the RSAD model to adopt a streamlined single-stream approach during inference. This streamlined architecture, comprised entirely of conventional neural network blocks, facilitates efficient processing and decision-making, allowing for swift anomaly detection in real-time industrial environments.

3.1. Stage 1: patch feature extractor

The Patch Feature Extractor operates by taking the original, untrimmed images as input and performing a series of operations to incorporate patch features into the Normalcy Library$L$, which is a compact normal feature space with clear boundaries. These operations involve analyzing the images at different levels of granularity to identify and capture local patch features that are essential for subsequent processing. Through this process, the Patch Feature Extractor effectively transforms the raw input images into a representation within Normalcy Library L that encapsulates the key characteristics of the underlying data, facilitating further analysis and interpretation downstream.

Specifically, given small-batch samples, we designate the training set and test set asχ_{tr} andχ_{te}, respectively. Initially, the Patch Feature Extractor employs a network ψ, pre-trained on ImageNet, for feature embedding, typically utilizing a ResNet-like backbone. Specifically, we define Z as the subset comprising the indexes of backbone layers. For each image $x_i \in R^{H\times W\times 3}$ in the set$\chi_{tr} \cup \chi_{te}$, the Patch Feature Extractor firstly extracts feature maps from the corresponding hierarchies $z \in Z$ denoted as

$$\psi^{z,i} \sim \psi^{z}(x_i) \in R^{H_z\times W_z\times C_z}, \tag{1}$$

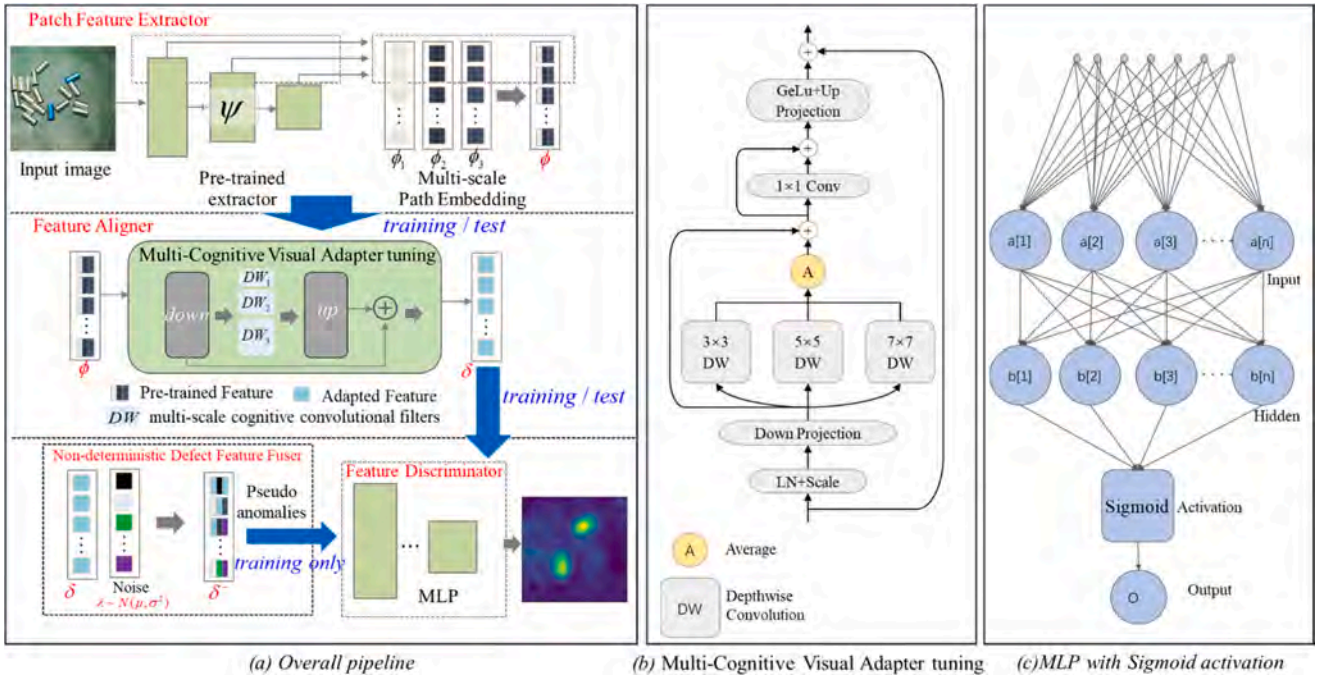

Fig. 1. (a) The overall pipeline of the proposed RSAD comprises four main stages. (b) The Multi-cognitive Visual Adapter tuning structure. (c) three-layer Multi-Layer Perceptron with Sigmoid activation.

Algorithm 1
RSAD training and testing pseudo-code.

1: **Input:** pretrained Patch Feature Extractor ψ, Feature AlignerA_γ, Defect Feature Fuser$N$, Feature Discriminator D_ε
2: **Training Stage:** $\psi, A_\gamma, N, D_\varepsilon$
3: Initialization:$N \leftarrow i.i.d.Gaussian$noise
$\gamma, \varepsilon \leftarrow randominitial$
4: Training sample $x_{train} \leftarrow normalsamples$
5: Pre-trained feature $f_{train} \leftarrow \psi(x_{train})$
6: Aligned feature $\phi_{train} \leftarrow A_\gamma(f_{train})$
7: Pseudo-anomalies $p_{anomaly} \leftarrow \phi_{train} + N$
8: Normal confidence $S_{AL}, S_{AL}^{-} \leftarrow D_\varepsilon\left(\phi_{train}, p_{anomaly}\right)$
9: $\mathscr{L} \leftarrow truncatedl_1 + focallossl_{foc}$
10: $\mathscr{L}.backward()$
11: $\gamma^*, \varepsilon^* \leftarrow \gamma, \varepsilon$ # update parameters of A and D
12: **Testing Stage:** $\psi, A_{\gamma^*}, D_{\varepsilon^*}$
13: Initialization:$A_{\gamma^*}, D_{\varepsilon^*} \leftarrow trainedweights_{\gamma^*, \varepsilon^*}$
14: test sample $x_{test} \leftarrow testdataset$
15: Pre-trained feature $f_{test} \leftarrow \psi(x_{test})$
16: Aligned feature $\phi_{test} \leftarrow A_{\gamma^*}(f_{test})$
17: Normal confidence $S_{AL} \leftarrow D_{\varepsilon^*}(\phi_{test})$
18: **Output:** anomaly heatmap S_{AL}

whereH_z,W_z and $C_z$represent the height, width, and channel size of the feature map $\psi^{z,i}$, respectively. Subsequently, we denote$\psi_{a,b}^{z,i} \in \mathbb{R}^{C_z}$as a C_z-dimensional feature slice positioned at (a, b), and its neighborhood with a patch size ∂ can be defined as

$$N_{a,b}^{\partial} = \{(a', b') | a' \in [a - \lfloor \partial / 2 \rfloor, ..., a + \lfloor \partial / 2 \rfloor]. \tag{2}$$

To prevent the extraction of overly generic or biased features that may be influenced by ImageNet classification, a specific strategy is employed within the Patch Feature Extractor. In this approach, the feature maps derived from hierarchy 4 are intentionally discarded, as they may not sufficiently capture the nuanced details relevant to the task at hand. Through meticulous analysis, it has been determined that the intermediate layers of the *WideResNet50* model, specifically the 2nd and 3rd layers, yield optimal results when configured with a neighborhood size of $\delta_{a,b}^{z}\partial = 3$. This strategic selection ensures that the extracted features are well-suited for subsequent processing, maximizing the effectiveness of the overall model.

Following the extraction of feature maps from each hierarchy, an adaptive average pooling mechanism is employed as the aggregation function, denoted asP_{agg}, to generate locally aggregated feature$\varphi_{a,b}^{z,i}$ in each level. Notably, this aggregation process, encompassing a local neighborhood, is pivotal for preserving essential spatial context within the feature maps. The process can be marked as

$$\varphi_{a,b}^{z,i} = P_{agg}\left(\psi_{a,b}^{z,i} \middle| (a', b') \in N_{a,b}^{\partial}\right). \tag{3}$$

This process serves as a crucial step in the overall Patch Feature Extractor, facilitating the extraction of relevant features for subsequent stages of processing.

Finally, to effectively integrate locally aggregated features into Normal Library$L$, a simple approach is adopted within the feature extraction pipeline. Initially, all feature maps undergo linear resizing to ensure uniformity in dimensions, with the size set to(a_0, b_0), corresponding to the dimensions of the largest feature map. Subsequently, these resized feature maps are concatenated channel-wise to generate the integrated patch feature map, denoted as$\phi^i \in \mathbb{R}^{a_0 \times b_0 \times c_z}$,where $\phi_{a,b}^i$, extracted patch feature at location (a, b) , is included in the Normalcy Library$L$. This integration process, represented as

$$\begin{aligned} \phi^i &= I_{cat}\left(resize\left(\varphi^{z',i}, (a_0, b_0)\right) \middle| z' \in Z\right) \\ L &= \bigcup_{x_i \in \chi_{tr}} \phi_{a,b}^i \end{aligned}, \tag{4}$$

facilitates the seamless combination of features across different levels of abstraction, thereby creating a comprehensive representation library that encapsulates relevant information from all hierarchical levels.

3.2. Stage 2: feature aligner

Adapter tuning refines pretrained vision models by introducing lightweight trainable modules while keeping the backbone frozen, ensuring efficient task-specific adaptation. In this work, we incorporate Feature Aligner $f_{ali}(\gamma)$ directly after the Feature Extractor, leveraging the Multi-Cognitive Visual Adapter tuning to bridge the domain gap between industrial images and natural image datasets like ImageNet. The learnable parameter γ enables the model to align pretrained features with industrial-specific characteristics while preserving the general knowledge of the pretrained network. This alignment process can be represented as:

$$\delta_{a,b}^i = f_{ali_\gamma}\left(\phi_{a,b}^i\right) \tag{5}$$

As illustrated in Fig. 1(b), the Mona-based Feature Aligner adopts a structured adaptation framework to refine feature representations while maintaining computational efficiency. The process begins with a scaled LayerNorm (LN) to normalize feature distributions, ensuring stability across varying domains. Subsequently, a down-projection layer reduces feature dimensionality, balancing representational capacity and computational cost. To enhance domain adaptability, we introduce multi-scale cognitive convolutional filters, leveraging depth-wise convolutions (DWConv) with kernel sizes of 3×3, 5×5, and 7×7. These filters extract spatial features at different receptive fields, capturing diverse contextual dependencies. The extracted features are then averaged and processed through a 1×1 convolutional layer for feature aggregation, followed by GeLU activation to introduce non-linearity and enhance expressiveness. To mitigate information loss and stabilize feature transformation, skip connections are incorporated at multiple stages. Finally, an up-projection layer restores the original feature dimensionality, ensuring seamless integration with downstream anomaly detection models.

By embedding Mona within the Feature Aligner, the proposed approach enables effective domain adaptation while preserving the efficiency and scalability of adapter tuning. This ensures robust feature alignment tailored to industrial anomaly detection scenarios.

3.3. Stage 3: non-deterministic defect feature fuser

To train the Feature Discriminator effectively, negative samples representing defect features are often synthesized since obtaining sufficient real anomalous samples in optimized industrial processes is challenging. However, synthetic anomalies may not accurately reflect

real defects, and the variability of actual anomalies limits the comprehensiveness of generated samples.

Thus, we introduce the Non-deterministic Defect Feature Fuser, which synthesizes negative samples by fusing simple noise onto normal samples within the feature space. Defect features are generated by fusing Gaussian noise with the normal features, denoted as $\delta^i_{a,b} \in \mathbb{R}^C$. Formally, a noise vector$\lambda \in \mathbb{R}^C$is sampled, with each entry following an independent identically distributed Gaussian distribution $N(\mu, \sigma^2)$. The scale of the noise, represented by σ, determines the extent to which the synthesized abnormal features deviate from the normal ones. We posit that by appropriately calibrating the scale of the noise, a tightly bounded normal feature space can be attained.

The resulting defect feature can be denoted as

$$\delta^{i-}_{a,b} = \delta^i_{a,b} + \lambda, \tag{6}$$

where$\delta^i_{a,b}$indicates the original normal feature. Non-deterministic Defect Feature Fuser provides a comprehensive representation that incorporates both normal and abnormal characteristics. It enables the Feature Discriminator to effectively discern between normal and anomalous samples, facilitating accurate anomaly detection in real-world industrial settings.

3.4. Stage 4: feature discriminator

Following the synthesis process in stage 3, the resulting defect feature$\delta^{i-}_{a,b}$and the normal aligned features $\{\delta^i | x_i \in \chi_{tr}\}$ are utilized as negative and positive samples, respectively, to train the final Feature Discriminator. This Feature DiscriminatorD_ε is designed to output positive values for normal features and negative values for anomalous features.

Drawing inspiration from [32], we design the Feature Discriminator as a three-layer MLP with a Sigmoid activation function, as shown in Fig. 1(c). Each layer consists of fully connected transformations, progressively refining feature representations. The final Sigmoid activation constrains the output within the range (0,1), providing an estimation of the normality of each location(a,b)through its output$D_\varepsilon(\delta_{a,b} \in \mathbb{R})$. For a given image$x_i \in \chi_{tr} \cup \chi_{te}$, the anomaly score for a feature at location (a, b) is represented ass$^i_{a,b}$ whereas the anomaly map can be denoted as$S_{AL}(x_i)$.

Considering that the most responsive point exists regardless of the size of the anomalous region, it is logical to evaluate the normalcy of samples by examining the maximum score on the anomaly map. Consequently, by providing a comprehensive measure of anomaly detection efficacy, the anomaly detection score $S_{AD}(x_i)$ for each image x_i can be calculated.

3.5. Loss function and optimizer

To enhance anomaly detection, we employ a combination of segmentation and classification losses. The segmentation loss refines anomaly localization, while the classification loss mitigates class imbalance and improves detection robustness.

For segmentation, we employ a truncated l_1 loss to restrict the influence of extreme values while preserving anomaly localization sensitivity. Specifically, the segmentation loss is computed as:

$$\mathscr{L}_{seg} = \sum_{x_i \in \chi_{tr}} \sum_{a,b} \frac{l^i_{a,b}}{a_0 * b_0},$$
$$l^i_{a,b} = \max\left(0, th^+ - D_\varepsilon\left(\delta^i_{a,b}\right)\right) + \max\left(0, -th^- + D_\varepsilon\left(\delta^{i-}_{a,b}\right)\right) \tag{7}$$

where th^+ and th^- are respectively set to 0.5 and −0.5 by default to prevent overfitting.

For anomaly detection, the Focal loss l_{foc} is employed to address the imbalance issue in binary classification. For each image $x_i \in \chi_{train}$ with ground truthM_{x_i}, the $\mathscr{L}_{cls}$is given by the Focal loss between the maximum score $S_{AD}(x_i)$and M_{x_i}. The classification loss is defined as:

$$\mathscr{L}_{cls} = \sum_{x_i \in \chi_{tr}} l_{foc}\left(S_{AD}(x_i), M_{x_i}\right) \tag{8}$$

The final loss is the sum of the segmentation and classification loss.

$$\mathscr{L} = \mathscr{L}_{seg} + \mathscr{L}_{cls}, \tag{9}$$

For updating the parameters of the Feature Aligner and Feature Discriminator, we employ the Adam optimizer. The initial learning rates are set to 0.0001 and 0.0002, respectively, with a weight decay of 0.00001.

4. Experiment setup

This section comprises a description of the dataset employed in the experiments, evaluation metrics, and comprehensive details of the experimental settings.

4.1. Datasets

MPDD: The Metal Parts Defect Detection dataset serves as a robust benchmark for evaluating anomaly detection methods for painted metal parts [6]. Curated to simulate real-world scenarios, it consists of 1346 images across six classes of metal parts, each meticulously labeled for accurate comparison of the proposed RSAD model with other algorithms. As shown in Table 1, the dataset is divided into nominal-only training data and test sets containing both normal and anomalous samples, with each sample accompanied by ground truth anomaly masks for precise assessment. The training set includes 888 normal samples, while the test set has 176 normal and 282 abnormal samples, ensuring a balanced representation of conditions in painted metal parts manufacturing.

Unlike other industrial anomaly detection datasets, the MPDD emphasizes variability, featuring samples with multiple objects in different positions and rotations against diverse backgrounds. Some samples also show components in motion, potentially introducing motion blur. This diversity captures a wide range of scenarios typical in the metal fabrication and painting industry, enhancing the dataset's relevance for real-world anomaly detection challenges.

Beyond MPDD, we evaluate our model on three industrial anomaly detection benchmarks: MVTec-AD [33], VisA [34], and PCB-Bank [13], each posing unique challenges. MVTec-AD consists of 10 object and 5 texture classes, with 3629 normal training and 1725 test images, including pixel-wise annotations for defects. VisA spans 12 categories with 9621 normal and 1200 anomalous high-resolution images, featuring complex structures, multiple objects, and diverse anomalies like scratches, dents, and structural defects. PCB-Bank integrates seven PCB categories, containing 4214 normal training and 2253 test samples, with anomalies including scratches, structural defects, and bending deformations, while also exhibiting variations in resolution, clarity, and viewing angles.

Table 1
The overall of MPDD dataset.

Class	Train	Test norm.	Test defect.
Bracket Black	289	32	47
Bracket Brown	185	26	51
Bracket White	110	30	30
Connector	128	30	14
Metal Plate	54	26	71
Tubes	122	32	69
Total	888	176	282

4.2. Evaluation metrics and baseline

In the anomaly detection task, the Feature Discriminator produces continuous outputs ($S_{AD}(x_i)$ and $S_{AL}(x_i)$)for image- level and pixel-level anomaly detection, respectively. To classify samples as normal or anomalous, a specific threshold is applied to these outputs. If $S_{AD}(x_i)$ is greater than the threshold ω, the image is labeled as normal; otherwise, it is classified as abnormal. Misclassifying a normal image as abnormal results in a false positive, indicating an erroneous decision.

Following established practices [35], AUROC (Area Under the ROC Curve) is selected as the primary metric for evaluating the RSAD model's performance. The ROC curve, plotting FPR against TPR, facilitates an assessment of the model's ability to discriminate between normal and anomalous samples. AUROC is particularly suitable for real-world applications, as it remains unaffected by class distributions, even in scenarios with class skew or significant changes.

For image-level evaluation (I-AUROC), $S_{AD}(x_i)$ is employed to compute AUROC, focusing on the overall accuracy of image classification. To achieve fine-grained anomaly localization, pixel-wise AUROC (P-AUROC) is calculated using $S_{AL}(x_i)$ to estimate the normality of each pixel and generate an anomaly map. In line with prior research, we compute on MPDD the class-average AUROC and mean AUROC across all categories for both detection and localization tasks.

The comparative baselines consist of recent SOTA methods, including reconstruction-based approaches (DMAD [11], DDAD [12], GLAD [13]), embedding-based techniques (MemKD [26], CFLOW [27], PatchCore [25]), and synthesizing-based models (Draem [20], RealNet [21], DiffusionAD [22]).

4.3. Experimental setup

All experiments were conducted using PyTorch on an NVIDIA GeForce GTX 2080Ti. Training was performed for 160 epochs with a batch size of 4. All images were resized to 256 × 256 and center-cropped to 224 × 224. The Adam optimizer was employed to train the Feature Aligner and Feature Discriminator, with initial learning rates of 0.0001 and 0.0002, respectively, and a weight decay of 0.00001. Our proposed RSAD framework comprises four stages. For feature extraction, we adopted WideResNet50 pretrained on ImageNet as the backbone ψ, extracting features from the 2nd and 3rd final outputs. The neighborhood patch size for feature aggregation was set to 3, and after reshaping and concatenation, the final feature dimension was 1536. For domain adaptation, we employed Multi-cognitive Visual Adapter tuning. Mona compresses pre-trained features into a low-dimensional space, with the intermediate dimension set to 64, ensuring optimal performance while maintaining efficiency. Subsequently, the Non-deterministic Defect Feature Fuser introduced independent Gaussian noise into normal feature embeddings. The scale of noise $\sigma$is set to 0.015 to achieve optimal performance. The final Feature Discriminator was implemented as a three-layer Multi-Layer Perceptron with Sigmoid activation.

5. Experiment result and analysis

The experiment results encompass two aspects: anomaly detection and anomaly localization.

5.1. Anomaly detection on MPDD

We conducted a comprehensive comparison of our proposed RSAD model with representative methods on the MPDD dataset, encompassing reconstruction-based, synthesizing-based and embedding-based methods. Table 2 presents the anomaly detection results on MPDD, where the image-level anomaly score is determined by the maximum score of the anomaly map, as defined as $S_{AD}(x_i)$. Notably, our RSAD model achieves a SOTA average I-AUROC of 98.3 %, outperforming previous methods and ranking first on 4 out of 6 classes. Especially noteworthy are the Connector and Metal Plate categories, where our model achieves perfect classification accuracy. In contrast, only an I-AUROC of 94.3 % is achieved for *Bracket Black* anomaly detection task, where defects contain scratches and holes.

5.2. Anomaly localization on MPDD

The anomaly localization performance is evaluated using pixel-wise AUROC (P-AUROC), as detailed in Table 3. Notably, our RSAD model achieves a SOTA average P-AUROC of 98.7 %, setting new benchmarks particularly for Connector and Tubes, with P-AUROC scores of 99.6 % and 99.2 %, respectively. Imperfectly, RSAD still seems relatively weaker at localizing bend and part defects on Bracket Brown, achieving a P-AUROC of 96.2 %, while performance on Bracket Black and Metal Plate remains strong, approaching 99.0 % (98.9 % and 98.8 %, respectively). To provide further insight, we illustrate representative samples for anomaly localization in Fig. 2.

5.3. Anomaly detection on other benchmarks

To evaluate the generalization of RSAD, we test it on three industrial anomaly detection benchmarks: MVTec-AD, VisA, and PCB-Bank, using image-level and pixel-level AUROC scores. The results are summarized in Table 4. On MVTec-AD, which includes both object and texture anomalies with pixel-wise annotations, RSAD achieves 96.1 % and 93.6 % AUROC, respectively, demonstrating strong performance in detecting and localizing industrial defects. On VisA, a high-resolution dataset with multi-object scenes and diverse anomaly types, RSAD achieves 93.8 % and 92.5 % AUROC, with a slight drop due to object occlusion and inter-class variations. On PCB-Bank, a dataset focused on printed circuit board defects with variations in resolution and viewing angles, RSAD achieves 97.4 % and 96.4 % AUROC, highlighting its effectiveness in structured industrial applications. Overall, RSAD attains an average AUROC of 95.8 % (image-level) and 94.2 % (pixel-level), confirming its robustness across diverse industrial scenarios.

Table 2
Comparison of image-level AUROC (I-AUROC%) for anomaly detection task on MPDD dataset. Best results are highlighted in bold.

Type	Reconstruction-based			Embedding-based			Synthesizing-based			Ours
Model	DMAD	DDAD	GLAD	MemKD	CFLOW	PatchCore	Draem	RealNet	DiffusionAD	RSAD
Bracket Black	80.5	**98.7**	98.0	95.7	72.7	81.9	91.8	94.9	97.5	94.3
Bracket Brown	94.5	92.7	90.7	98.9	88.8	78.4	90.3	96.8	93.8	**99.7**
Bracket White	82.9	96.6	**98.3**	98.3	87.8	76.0	88.8	88.8	88.7	96.1
Connector	99.0	96.2	**100.0**	**100.0**	94.8	96.7	100.0	**100.0**	97.4	**100**
Metal Plate	**100.0**	**100.0**	99.9	**100.0**	99.5	**100.0**	**100.0**	**100.0**	**100.0**	**100**
Tubes	93.4	99.2	98.1	95.6	73.1	59.7	94.7	97.5	99.7	**99.7**
Avg.	91.7	97.2	97.5	98.1	86.1	82.1	94.3	96.4	96.2	**98.3**

Table 3
Comparison of pixel-level AUROC (P-AUROC%) for anomaly localization task on MPDD dataset. Best results are highlighted in bold.

Type	Reconstruction-based			Embedding-based			Synthesizing-based			Ours
Model	DMAD	DDAD	GLAD	MemKD	CFLOW	PatchCore	Draem	RealNet	DiffusionAD	RSAD
Bracket Black	91.1	96.7	**99.4**	97.8	96.9	98.4	98.2	99.3	98.3	98.9
Bracket Brown	81.6	97.2	97.5	96.3	**97.8**	91.5	63.7	**97.8**	93.1	96.2
Bracket White	93.2	91.8	**99.7**	98.8	98.6	97.4	98.9	97.4	93.5	99.4
Connector	97.7	98.6	98.2	99.4	98.4	95.0	91.2	97.5	94.5	**99.6**
Metal Plate	95.6	98.1	**99.4**	99.1	98.2	96.6	96.6	99.3	94.7	98.8
Tubes	96.5	99.0	97.8	**99.2**	96.4	95.1	95.9	97.9	97.8	**99.2**
Avg.	92.6	96.9	**98.7**	98.4	97.7	95.7	90.7	98.2	95.3	**98.7**

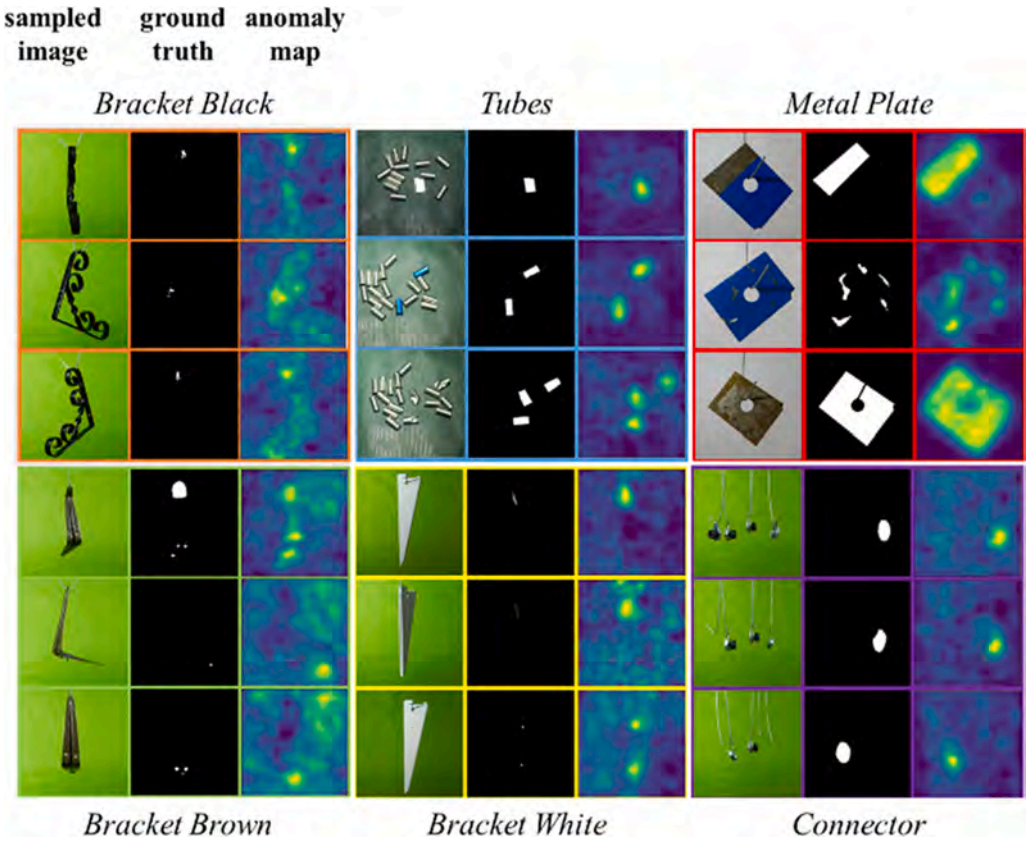

Fig. 2. Qualitative results, where sampled image (left), ground truth (middle), and anomaly map (right) are shown for each class in MPDD dataset.

Table 4
Generalization performance of RSAD across industrial benchmarks (Image-level and Pixel-level AUROC %).

Benchmarks	MVTec-AD	Visa	PCB-Bank	Avg
I- AUROC	96.1	93.8	97.4	95.8
P-AUROC	93.6	92.5	96.4	94.2

6. Ablation studies on feature aligner

To validate the effectiveness of the Feature Aligner in anomaly detection, we designed an ablation study comparing different alignment strategies. The Feature Aligner essentially functions as adapter tuning, incorporating a lightweight trainable module that enables effective domain adaptation without modifying the core parameters. Our pre-trained backbone, WideResNet-50, consists of 25.5 million parameters, and full fine-tuning the model for each task would be computationally expensive and prevent the utilization of pretrained general knowledge. In contrast, Feature Aligner preserves the efficiency and scalability of the pre-trained network, as well as adapting pretrained features with industrial-specific characteristics

Without performing full fine-tuning, we designed three experimental setups: (1) a baseline without an adapter, (2) tuning with a bias-free fully connected (FC) layer, and (3) tuning with Mona. To evaluate the trade-off between performance improvement and parameter efficiency, we analyzed the additional parameter overhead introduced by each approach. For the FC Layer, with both input and output dimensions set to 1536, the additional parameters amount to 2.36 million (1536×1536). For Mona, with an input dimension of $m = 1536$ and an intermediate down-projected dimension of $n = 64$, the total trainable parameters are [4]:

$$2 \times ((2n+3)m + n2 + 84n + 2) = 0.42\text{million} \tag{10}$$

Both approaches introduce minimal parameter overhead relative to the pretrained WideResNet-50 backbone (25.5 M parameters), particularly Mona, which is significantly more lightweight.

To assess the impact of adapter tuning, we compare the anomaly detection performance using I-AUROC and P-AUROC on the MPDD dataset. As shown in Table 5, the FC Layer improves I-AUROC by 2.3 % and P-AUROC by 3.2 %, but requires 2.36 M additional parameters (9.2

Table 5
Comparison of Feature Aligner configurations in terms of detection performance and additional parameter overhead.

Methods	Parameters /backbone	I-auroc(%)	p-auroc(%)
Mona (Ours)	1.65 %	98.3	98.7
No Adapter	/	95.3	94.5
FC Layer	9.2 %	97.6	97.7

% of the backbone parameters). Mona introduces only 0.42 M additional parameters (1.65 % of the backbone) yet achieves further gains of +0.7 % I-AUROC and +1.0 % P-AUROC over the FC Layer, using just 17.9 % of its parameter count.

These findings demonstrate that Mona provides a superior balance between anomaly detection accuracy and parameter efficiency, validating adapter tuning as an effective domain adaptation strategy for industrial anomaly detection.

7. Conclusion

The proposed RSAD model is a four-stage unsupervised anomaly detection framework designed to address real-world challenges in mechanical manufacturing. It detects random surface defects without requiring labeled anomaly samples and adapts to data distribution shifts in complex industrial environments, ensuring consistent product quality.

At the heart of the RSAD model lies a series of simple neural network modules designed to extract locally aggregated, mid-level features from the 2nd and 3rd final outputs of a WideResNet50 model pretrained on ImageNet. Subsequently, we incorporate a Feature Aligner directly after the Feature Extractor, leveraging the Multi-Cognitive Visual Adapter tuning to bridge the domain gap between industrial images and pretrained datasets. This approach adds only 1.65 % of the backbone's parameters yet improves I-AUROC by 3.0 % and P-AUROC by 4.2 %. In the third stage, synthesizing-based methods fuse Gaussian noise with normal features in the feature space, generating non-deterministic, anomalous samples. Finally, Feature Discriminator serves as a normality scorer, estimating the likelihood of samples being normal.

Extensive experiments on the MPDD dataset and other representative benchmarks demonstrate the RSAD model's SOTA performance in anomaly detection and localization. This model is a significant advancement in visual intelligence for quality inspection across diverse product surfaces, enhancing the reliability of quality control processes in mechanical manufacturing.

CRediT authorship contribution statement

Pulin Li: Writing – review & editing, Supervision, Project administration, Funding acquisition, Formal analysis, Conceptualization. **Guocheng Wu:** Writing – original draft, Visualization, Validation, Software, Methodology, Conceptualization. **Yanjie Zhou:** Writing – review & editing, Validation, Funding acquisition. **Jiewu Leng:** Writing – review & editing, Validation, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by the National Natural Science Foundation of China with Grant 52305559, the China Postdoctoral Science Foundation with Grant 2023M733208, and the Humanities and Social Science Project of Henan with Grant 2024-ZZJH-012. (Corresponding author: Yanjie Zhou).

Data availability

Data will be made available on request.

References

[1] X. Xiang, M. Liu, S. Zhang, P. Wei, B. Chen, Multi-scale attention and dilation network for small defect detection, Pattern Recognit. Lett. 172 (2023) 82–88.
[2] Y. Xia, Y. Lu, X. Jiang, M. Xu, Enhanced multiscale attentional feature fusion model for defect detection on steel surfaces, Pattern Recognit. Lett. 188 (2025) 15–21.
[3] V. Zavrtanik, M. Kristan, D. Skočaj, Keep dræming: discriminative 3D anomaly detection through anomaly simulation, Pattern Recognit. Lett. 181 (2024) 113–119.
[4] D. Yin, L. Hu, B. Li, Y. Zhang, X. Yang, 5%>100%: breaking performance shackles of full fine-tuning on visual recognition tasks, arXiv e-Prints (2024) arXiv: 2408.08345.
[5] S. Deng, et al., EHIR: energy-based hierarchical iterative image registration for accurate PCB defect detection, Pattern Recognit. Lett. 185 (2024) 38–44.
[6] S. Jezek, M. Jonak, R. Burget, P. Dvorak, M. Skotak, Deep learning-based defect detection of metal parts: evaluating current methods in complex conditions, in: Proceedings of the IEEE 13th International Congress on Ultra Modern Telecommunications and Control Systems, Brno, Czech Republic, 2021, pp. 66–71.
[7] M. Kwon, Y. Moon, B. Lee, J. Noh, Autoencoders with exponential deviation loss for weakly supervised anomaly detection, Pattern Recognit. Lett. 171 (2023) 131–137.
[8] S. Akçay, A. Atapour-Abarghouei, T.P. Breckon, Skip-GANomaly: skip connected and adversarially trained encoder-decoder anomaly detection, in: Proceedings of the International Joint Conference on Neural Networks, Budapest, Hungary, 2019, pp. 1–8.
[9] S. Akcay, A. Atapour-Abarghouei, T.P. Breckon, GANomaly: semi-supervised anomaly detection via adversarial training, in: Proceedings of the Asian Conference on Computer Vision, Springer, Cham, 2019, pp. 622–637.
[10] H. He, et al., MambaAD: exploring State space models for multi-class unsupervised anomaly detection. Advances in Neural Information Processing Systems, 2024, pp. 71162–71187.
[11] W. Liu, H. Chang, B. Ma, S. Shan, X. Chen, Diversity-measurable anomaly detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 12147–12156.
[12] A. Mousakhan, T. Brox, J. Tayyub, Anomaly detection with conditioned denoising diffusion models, arXiv e-Prints (2023) arXiv:2305.15956.
[13] H. Yao, M. Liu, Z. Yin, Z. Yan, X. Hong, W. Zuo, GLAD: towards better reconstruction with global and local adaptive diffusion models for unsupervised anomaly detection, in: Proceedings of the European Conference on Computer Vision, Springer Nature Switzerland, Cham, 2025, pp. 1–17.
[14] Z. You, et al., A unified model for multi-class anomaly detection. Advances in Neural Information Processing Systems, Curran Associates, 2022, pp. 4571–4584.
[15] H. He, et al., A diffusion-based framework for multi-class anomaly detection, in: Proceedings of the AAAI Conference on Artificial Intelligence 38, 2024, pp. 8472–8480.
[16] J. Zhang, et al., Exploring plain ViT reconstruction for multi-class unsupervised anomaly detection, arXiv e-Prints (2024) arXiv:2312.07495.
[17] J. Zhang, et al., Learning feature inversion for multi-class anomaly detection under general-purpose COCO-AD benchmark, arXiv e-Prints (2024) arXiv:2404.10760.
[18] C. Li, K. Sohn, J. Yoon, T. Pfister, CutPaste: self-supervised learning for anomaly detection and localization, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 2021, pp. 9659–9669.
[19] J. Fan, Y. Ji, H. Wu, Y. Ge, D. Sun, J. Wu, An unsupervised video anomaly detection method via Optical Flow decomposition and Spatio-temporal feature learning, Pattern Recognit. Lett. 185 (2024) 239–246.
[20] V. Zavrtanik, M. Kristan, D. Skočaj, DRÆM – A discriminatively trained reconstruction embedding for surface anomaly detection, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, Canada, 2021, pp. 8310–8319.
[21] X. Zhang, M. Xu, X. Zhou, RealNet: a feature selection network with realistic synthetic anomaly for anomaly detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 16699–16708.
[22] H. Zhang, Z. Wang, Z. Wu, Y.-G. Jiang, DiffusionAD: norm-guided one-step denoising diffusion for anomaly detection, arXiv e-Prints (2023) arXiv: 2303.08730.
[23] T. Defard, A. Setkov, A. Loesch, R. Audigier, PaDiM: a patch distribution modeling framework for anomaly detection and localization, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, Virtual, Milan, Italy, 2021, pp. 475–489.
[24] K. Zhang, Y.P. Tsang, C.K.M. Lee, C.H. Wu, Integrating large language models with explainable fuzzy inference systems for trusty steel defect detection, Pattern Recognit. Lett. 192 (2025) 29–35.
[25] K. Roth, L. Pemula, J. Zepeda, B. Scholkopf, T. Brox, P. Gehler, Towards total recall in industrial anomaly detection, in: Proceedings of the IEEE/CVF Conference on

Computer Vision and Pattern Recognition, New Orleans, LA, USA, IEEE, 2022, pp. 14298–14308.
[26] Z. Gu, et al., Remembering normality: memory-guided knowledge distillation for unsupervised anomaly detection, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 16401–16409.
[27] D. Gudovskiy, S. Ishizaka, K. Kozuka, CFLOW-AD: real-time unsupervised anomaly detection with localization via conditional normalizing flows, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 2022, pp. 1819–1828.
[28] R. He, et al., On the effectiveness of adapter-based tuning for pretrained language model adaptation, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing 1, 2021, pp. 2208–2222. Long Papers.
[29] S. Chen, et al., AdaptFormer: adapting vision transformers for scalable visual recognition. Advances in Neural Information Processing Systems, Curran Associates, Inc, 2022, pp. 16664–16678.
[30] X. He, C. Li, P. Zhang, J. Yang, X.E. Wang, Parameter-efficient model adaptation for vision transformers, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2023, pp. 817–825.
[31] D. Yin, Y. Yang, Z. Wang, H. Yu, K. Wei, X. Sun, 1% vs 100%: parameter-efficient low rank adapter for dense predictions, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2023, pp. 20116–20126.
[32] Z. Liu, Y. Zhou, Y. Xu, Z. Wang, Simplenet: a simple network for image anomaly detection and localization, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 2023, pp. 20402–20411.
[33] P. Bergmann, M. Fauser, D. Sattlegger, C. Steger, MVTec AD — A comprehensive real-world dataset for unsupervised anomaly detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 2019, pp. 9584–9592.
[34] Y. Zou, J. Jeong, L. Pemula, D. Zhang, O. Dabeer, SPot-the-difference self-supervised pre-training for anomaly detection and segmentation, in: Proceedings of the European Conference on Computer Vision, Springer Nature Switzerland, Cham, 2022, pp. 392–408.
[35] J. Yu, H. Oh, Y. Lee, J. Yang, Denoising diffusion model with adversarial learning for unsupervised anomaly detection on brain MRI images, Pattern Recognit. Lett. 186 (2024) 229–235.