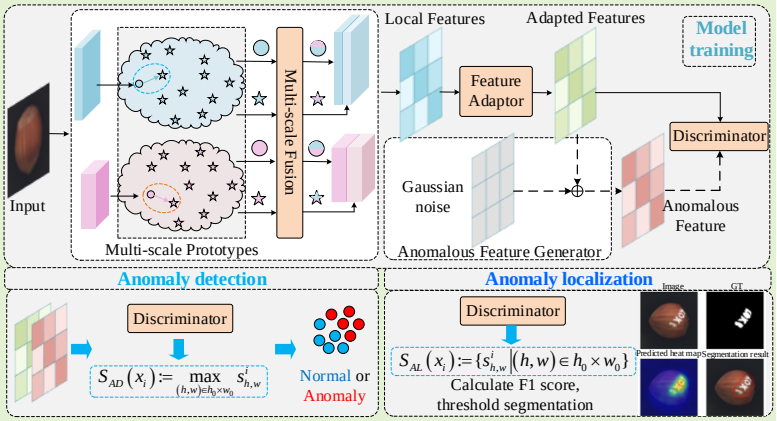
Multi-scale Prototype Fusion Network for Industrial Product Surface Anomaly Detection and Localization

Haidong Shao, *Senior Member, IEEE*, Jiangji Peng, Minghui Shao, and Bin Liu, *Member,*

***Abstract*—In complex industrial application scenarios, abnormal samples are scarce. In the case of weak defect features, the high similarity of positive and negative samples further complicates detection and localization. In addition, anomalies are often subtle and unpredictable, which makes it particularly difficult to detect and localize anomalous subregions with unknown anomaly patterns. Many detection algorithms suffer from high computational complexity and huge memory consumption. To address these challenges, this paper proposes a multi-scale prototype fusion network for industrial product surface anomaly detection and localization (MPFnet). MPFnet uses multi-scale prototypes to construct representative normal patterns and incorporates a multi-scale fusion block to facilitate information exchange between different scales. This design enhances the model's attention to characterize prototype and normal features. Feature adapter is constructed to generate fitness features, reducing domain bias. By adding noise to the adapted features, anomalous features are generated, and anomalies are detected using a simple and efficient discriminator. A large number of experiments were carried out on the challenging MVTec AD and MVTec LOCO AD datasets, demonstrating that MPFnet outperforms other state-of-the-art comparative methods, achieving good detection and localization results regardless of defect patterns.**



***Index Terms*—anomaly detection and localization; prototype fusion network; multi-scale information exchange; industrial application scenarios; feature adapter;**

I. Introduction

In the field of industrial product surface inspection, the anomaly detection and localization task focuses on detecting anomalous images and accurately locating anomalous subregions, which is critical for ensuring product quality and production process stability. Despite extensive research in this area, various visible and invisible challenges persist [1]-[3]. In industrial application scenarios, anomalous samples are scarce and significantly outnumbered by normal samples, leading to a natural unbalanced learning problem [4]. Anomalies are usually subtle, ranging from subtle changes to large structural defects, and the presentation of anomalies is unpredictable, making it difficult for accurate localization [5]. As shown in Fig. 1(a), the surfaces of industrial products often exhibit complex and irregular background textures, complicating the recognition of defective samples. As shown in Fig. 1(b), the types of defects that occur during the manufacturing process are often unpredictable, and a range of different production processes coupled with inherent uncertainty can lead to differences between defect samples. In addition, as shown in Fig. 1(c), the high degree of similarity between defect-free and defective samples adds to the difficulties in identifying normal and abnormal samples, especially when defect characteristics are weak.

Abnormal samples are scarce in industrial scenarios, and supervised methods are difficult to effectively achieve the abnormality detection and localization tasks. Currently unsupervised learning has become the mainstream method, which uses only normal samples in the training process, and is primarily divided into three categories, which are reconstruction-based methods, synthesis-based methods, and embedding-based methods. Reconstruction-based approaches argue that if the model is trained with normal samples only, it will not be able to reconstruct anomalous regions efficiently.

This research is supported by the National Natural Science Foundation of China (No. 52275104), and the Science and Technology Innovation Program of Hunan Province (No. 2023RC3097). *(Corresponding author: Haidong Shao.)*

Haidong Shao, Jiangji Peng, and Minghui Shao are with College of Mechanical and Vehicle Engineering, Hunan University, Changsha 410082, China (e-mail: hdshao@hnu.edu.cn; pjj181@163.com; hui11@hnu.edu.cn).

Bin Liu is with Department of Management Science, University of Strathclyde, Glasgow G1 1XG, UK (email: b.liu@strath.ac.uk).

Anomaly detection is accomplished by evaluating the reconstruction error against the potential error. However, sometimes the models unpredictably reconstruct the abnormal samples well, which leads to accuracy degradation [6]. In addition, this method relies on the quality of normal samples, performs poorly on uncalibrated or noisy datasets, and its robustness needs to be improved [7]. The synthetic-based approach effectively gets rid of the problem of data imbalance by generating abnormal data on normal samples for training, thus constructing the decision boundary between normal and abnormal. However, this approach is opaque to true anomalies, and the features of the synthetic data may be far from the normal features, leading to loosely bounded normal feature space.

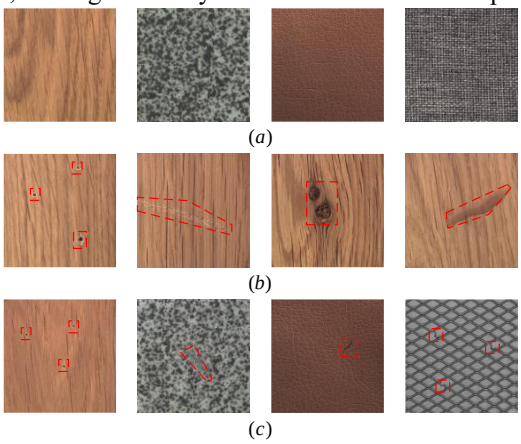


Fig. 1. (a) Complex and irregular background textures. (b) Diverse defect types. (c) Weak defect features.

The embedding-based approach using a convolutional neural network (CNN) pre-trained on ImageNet [8] is used to extract the generalized normal features of the image. Then, the distribution of normal features is characterized by statistical methods such as multivariate Gaussian models, normalized flow techniques, and memory banks. Anomaly detection is performed by checking the distributional differences between the input features and what has been mastered or memorized. However, the distributional properties of industrial images may be different from the pre-trained dataset of ImageNet, which may lead to mismatches when these features are used directly. At the same time, these statistical methods may encounter challenges of high computational complexity or memory usage in practice [9].

To address the above challenges, this paper proposes a multi-scale prototype fusion network (MPFNet) for industrial product surface anomaly detection and localization. Firstly, we propose multi-scale prototypes to capture and represent normal patterns. Unlike other methods that construct normal patterns by connecting feature memories or randomly sampling feature maps [10], we use intermediate feature map prototypes at different scales to construct normal patterns, providing accurate and representative normal mode characterization while maintaining the integrity of spatial information. On this basis, multi-scale fusion block are added for information exchange between different scales, aiming to further enhance the orthonormal sample utilization. Secondly, in place of directly applying features obtained from pre-training, we've designed a feature adapter to produce domain-specific features, aiding in the reduction of biases that arise from domain discrepancies. Subsequently, under the premise of ensuring that the normal feature space has compact boundaries, by adjusting the noise scale, we adopt a strategy of adding noise to normal features in the feature space as a way to construct anomalous features. Our method shows significant savings in model resource consumption compared to other methods that employ synthetic noise. Finally, we simplify the anomaly detection process and improve the computational efficiency by training a simple discriminator compared to the complex statistical algorithm employed by most embedding-based methods. The main contributions of this paper are summarized as follows:

(1) Multi-scale prototype fusion network for industrial product surface anomaly detection and localization is proposed. Representative normal patterns are constructed using intermediate feature map prototypes at different scales. Adding multi-scale fusion blocks to realize information exchange between different scales makes the model focus more on the prototype features and normal features.

(2) A feature adapter is constructed to generate target-oriented features in order to reduce the distribution bias between pre-trained and target features.

(3) Gaussian noise is added to normal features in the feature space to generate anomalous features, and a simple and efficient discriminator is trained to simplify the process of anomaly detection, which in turn improves computational efficiency.

This paper is structured as follows: Section 2 reviews related studies. Section 3 details the proposed approach. Section 4 outlines the anomaly detection and localization framework. Section 5 assesses the method's benefits and effectiveness through experiments. Section 6 concludes the paper.

II. Related Works

Anomaly detection and localization techniques fall into three main categories: reconstruction-based methods, synthesis-based methods, and embedding-based methods.

Reconstruction-based methods train models exclusively on normal samples so that the models learn to reconstruct the distributional features of these normal samples accurately. These types of methods operate under the assumption that the parameters of the model are obtained by training on normal samples, and that the model can reconstruct normal samples well but will struggle to accurately reconstruct the defective regions of anomalous samples. Defect detection and localization are then achieved by analyzing the reconstruction errors, which highlight the discrepancies between the original and reconstructed images. However, the performance of these methods heavily depends on the quality of the normal samples. Consequently, they tend to perform poorly on uncalibrated or noisy datasets, where the reconstruction errors may not accurately indicate anomalies. The AE-based method uses an encoder-decoder network, where the encoder encodes the input image as a hidden space variable z and the decoder reconstructs the image using the hidden space variable z. Anomaly detection and the pinpointing of defect are accomplished by evaluating the discrepancy between the original input image and its reconstructed image. However, the image reconstructed by AE is blurred, which can easily cause

false detection of the image with pixel-level accuracy when calculating the reconstruction error. To improve the detection performance, Zhou *et al*. [11] synthesized the information of image space and hidden space for defect detection on fabric surface. They combined VAE and Gaussian mixture model (GMM), used VAE to extract the features of the input image and reconstructed the image, and used structural similarity SSIM [12] to measure the reconstruction error. Generative modeling has been widely introduced for better reconstruction performance. GANomaly [13] adds adversarial training to AE by adding an additional encoder after the generator, which is used to constrain the reconstructed image to keep the same high-level features as the original image. In order to improve the quality of reconstruction, Skip-GANomaly [14] introduces skip-connections to fully incorporate features from normal samples at multiple scales, which improves the detection capability compared to GANomaly, but the performance varies widely on different classes of data. Ristea *et al*. [15] propose a self-supervised predictive convolutional attention block (SSPCAB), which is implemented by predicting the occluded regions in the convolutional receptive field and is trained by minimizing the reconstruction error of the occluded regions. However, the assumptions of the image reconstruction-based approach are not entirely reliable, and anomalies in an image can also be well reconstructed if they share a common compositional pattern with normal training data, or if the decoder is overfitted, leading to false detections [16].

Synthesis-based methods usually synthesize anomalies on normal samples. Li *et al*. [17] proposed CutPaste, which randomly crops and pastes other regions of a normal image into normal samples to obtain anomalous samples with pixel-level annotations, which are then utilized to learn representations of the model on a classification task. For testing, Gaussian density estimation is performed in the image and image block dimensions for defect detection and localization, respectively. Song *et al*. [18] added data enhancement to the source image with random rotation, positional disruption, and color dithering to improve the diversity of the synthetic defects on top of CutPaste, and spliced coordinate channels to improve the segmentation performance, using two branches to perform the reconstruction of the normal region and the anomalous region segmentation. DRAEM [19] proposes an end-to-end network architecture dedicated to discriminative training of synthesized generated just-out-of-distribution patterns. However, the synthesized anomalies do not always match the real anomaly patterns. In addition, synthesis-based methods are very sensitive to the way they are synthesized and often tend to overfit to artificially defective patterns.

Embedding-based methods compress and embed normal features into a dense space, while anomalous features behave as outliers that are significantly distant from the normal clusters. Most of such methods utilize pre-trained networks on ImageNet for feature extraction. PaDiM [20] utilizes a pre-trained model for patch embedding and uses a multivariate Gaussian distribution to obtain a probabilistic representation of the normal category, which localizes the anomalies using correlations between different semantic hierarchies. PatchCore [21] uses a maximal representational memory bank containing the normal patch features. The input features are scored in tests using either the Mahalanobis distance or the maximum feature distance. However, industrial images often have different distributions than ImageNet, and using pre-trained features directly may lead to mismatch problems [9]. CS-Flow [22] uses normalized flow to transform normal feature distributions into Gaussian distributions by jointly processing multiple feature maps at different scales. DRA [23] proposes to achieve anomaly detection by learning a detection model using labeled anomaly samples, allowing the model to capture decoupled anomaly representations described by known anomalies, pseudo-anomalies, and potential residual anomalies. CFLOW [24] uses the Conditional Normalized Flow framework to learn the probability distribution of feature vectors by transforming an easy-to-handle base distribution to fit an arbitrary target distribution. FastFlow [32] provides a novel unsupervised anomaly detection and localization method by considering global and local information as well as a learnable distribution modeling approach and an efficient inference process. However, such methods are memory intensive since normalized flow does not allow down-sampling and the coupling layer consumes several times more memory than a normal convolutional layer.

The proposed MPFNet can effectively overcome the above issues. MPFNet learns multi-scale feature maps at each scale and uses multi-scale fusion block to realize information exchange between different scales, which makes the model focus on the prototypical and normal features in a more detailed way. Additionally, feature adapter are constructed to generate target-oriented features to reduce domain bias to accurately detect and localize anomalous samples of various patterns.

III. The Proposed Method

In this section, we provide a detailed description of the proposed MPFNet, depicted in Fig. 2. It comprises a multi-scale prototype fusion feature extractor, a feature adapter, an anomaly feature generator, and a discriminator. Among them, the multi-scale prototype fusion feature extractor consists of multi-scale prototypes and multi-scale fusion block, etc., and the anomaly feature generator is used only for training.

1. Multi-scale Prototype Fusion Feature Extractor

A. Prototype Initialization

Denote the training set and the test set as χ_{train} and χ_{test}, respectively. Input image $x_i \in \mathbb{R}^{H\times W\times 3}$ in the training set and test set. We harness a pre-trained ImageNet network to extract multi-scale feature representations of the input images. The feature mapping $F_{i,j} = F_j(x_i)(j \in \{2,3\})$ represents the output of input x_i at the j-th block of the pre-trained network, where $F_{i,j} \in \mathbb{R}^{c^j \times h^j \times w^j}$ is a tensor with depth c^j, height h^j, and width w^j. The j-th scale prototype $P_j \in \mathbb{R}^{K\times c^j \times h^j \times w^j}$ is a randomly selected K feature mapping from $F_j(\chi_{train})$, updated using k-means clustering [25]. Two scales of prototypes were used $(j \in \{2,3\})$. The model parameters were kept constant throughout the clustering process. Post-clustering, the prototypes $P_j \in \mathbb{R}^{K\times c^j \times h^j \times w^j}$ at each tier remain static during subsequent model training.

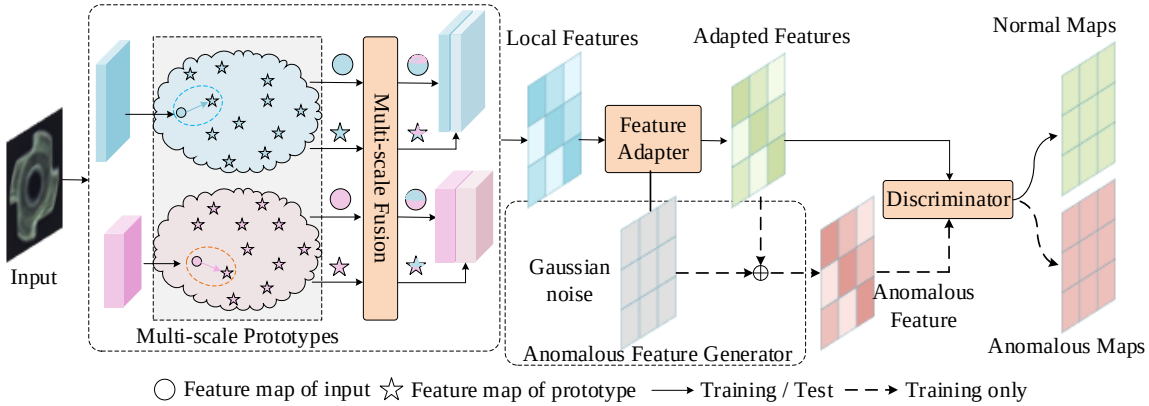


Fig. 2. MPFNet overall structure.

B. Multi-scale Fusion

In order to realize the exchange of information between different scales and make the model pay more attention to the various features of the input in a more detailed way [26], we built a multi-scale fusion module, as shown in Fig. 3. The fused feature mapping $F_{i,j}^{*}$ is as follows:

$$F_{i,j}^{*} = f_{2j}\left(F_{i,2}\right) + f_{3j}\left(F_{i,3}\right) \tag{1}$$

where the choice of the transformation function $f_{rj}(\bullet)$ depends on the input feature mapping index r and the output feature mapping index j, $(r, j \in \{2,3\})$. When $r = j$, then $F_{i,r} = f_{rj}\left(F_{i,r}\right)$. When $r < j$, $f_{rj}\left(F_{i,r}\right)$ down-samples the input feature mapping $F_{i,r}$ by a deeply separable convolution with step size 2^{j-r}, kernel size $2^{j-r}+1$, and padding 2^{j-r-1}. When $r > j$, $f_{rj}\left(F_{i,r}\right)$ up-samples the input feature mapping $F_{i,r}$ by bilinear up-sampling and 1×1 convolution. Similarly, the input features are processed identically and concatenated with $F_{i,j}^{*}$ along the depth dimension to obtain the multi-scale fused feature $C_{i,j}^{*} \in \mathbb{R}^{2c^{j}\times h^{j}\times w^{j}}$.

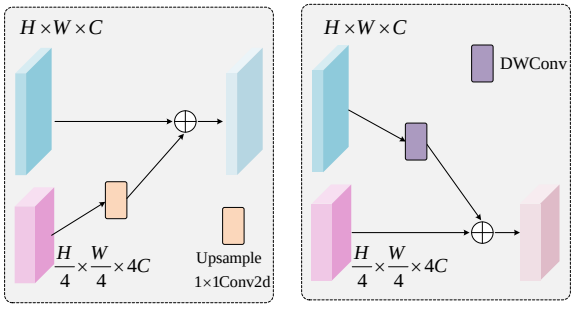


Fig. 3. Multi-scale fusion block fuses feature maps at different scales.

In order to retain enough detail information of the multi-scale fused features while not allowing the extracted abstract information to be too biased towards ImageNet data, feature aggregation is performed on local neighborhoods to extract the features, and for the feature $C_{i,j}^{*} \in \mathbb{R}^{2c^{j}\times h^{j}\times w^{j}}$ at position (h,w) after multi-scale fusion, the neighborhood with a patchsize of p [20] is defined as follows:

$$N_{p}^{(h,w)} = \{(h',y') \mid h' \in [h-[p/2],\cdots,h+[p/2]], \; y' \in [w-[p/2],\cdots,w+[p/2]]\} \tag{2}$$

where $N_{p}^{(h,w)}$ is a piece of patch of size $p\times p$ at position (h,w) on the feature map, and in this paper we take $p=3$.

The features within the neighborhood $N_{p}^{(h,w)}$ are aggregated using adaptive mean pooling to obtain the local feature $z_{h,w}^{*}$:

$$z_{h,w}^{*} = f_{agg}\left(\{C^{*}(h',y') \mid (h',y') \in N_{p}^{(h,w)}\}\right) \tag{3}$$

In order to combine features from different levels $z_{h,w}^{*}$, all feature mappings are linearly scaled to the same size (h_0, w_0), and feature mappings are connected by channel to obtain a feature mapping $o^{i} \in \mathbb{R}^{h_0\times w_0\times c}$:

$$o^{i} = f_{cat}\left(resize\left(z'_{h,w},(h_0,w_0)\right) \mid z'_{h,w} \in z_{h,w}^{*}\right) \tag{4}$$

2. Feature Adapter and Anomalous Feature Generator

Industrial images often have different data distributions than ImageNet pre-trained models, and directly using pre-trained features may lead to mismatch issues. Therefore, instead of directly utilizing the features of the pre-trained model, we reduce the inter-domain bias by adapting the features to the target domain using a feature adapter consisting of a single fully connected layer. Such feature adapter have been shown to achieve efficient performance [9]. Where the input and output features of the fully connected layer have the same dimensions. The feature adapter G_{θ} projects the local features $o_{h,w}^{i}$ to the adapted features:

$$q_{h,w}^{i} = G_{\theta}\left(o_{h,w}^{i}\right) \tag{5}$$

In anomaly detection techniques, a large number of synthetic strategies are used to train models by artificially generating anomalous samples on normal samples, with the aim of delineating the decision boundary between normal and anomalous samples. Such an approach is helpful to alleviate the imbalance between the number of normal and abnormal samples in the dataset.

However, this approach is opaque to true anomalies, and the features of the synthetic data may be far from the normal features, resulting in a loosely bounded normal feature space. We obtain tightly bounded normal feature space by appropriately calibrating the scale of the noise. In this paper, Gaussian noise is added to the normal sample feature space to generate anomalous features $q_{h,w}^{i-}$. The Gaussian noise satisfies the distribution $N(0,\sigma^2)$, where σ is set to 0.018. The anomalous features are defined as:

$$q_{h,w}^{i-} = q_{h,w}^{i} + \varepsilon \tag{6}$$

3. Discriminator and Loss Function

The discriminator consists of a linear layer, a batch normalization layer, a leaky relu (0.2 slope) layer, and a linear layer, which directly estimates the normality of each position (h,w). The discriminator is trained with a batch of anomalous features and a batch of anomalous features. The normal feature $\{q_{h,w}^{i} | x_i \in \chi_{train}\}$ and the abnormal feature $q_{h,w}^{i-}$ are fed into the discriminator for training together. The discriminator expects the output of the normal features to be positive and the output of the abnormal features to be negative, estimating the normality to be $D(q_{h,w}) \in \mathbb{R}$.

The discriminator is expected to judge the output of normal features as positive and the output of abnormal features as negative as possible. Therefore, we design the truncated loss $l1$. The truncated loss $l1$ is defined as:

$$l_{h,w}^{i} = \max\left(0, th^{-} - D\left(q_{h,w}^{i}\right)\right) + \max\left(0, -th^{-} + D\left(q_{h,w}^{i-}\right)\right) \tag{7}$$

where th and th^{-} are truncation terms to prevent overfitting. The final training objective is:

$$L = \min_{\theta} \sum_{x_i \in \chi_{train}} \sum_{h,w} \frac{l_{h,w}^{i}}{h_0 * w_0} \tag{8}$$

4. Abnormal Scoring Function

For inference, the test image $x_i \in \chi_{test}$ is input into MPFNet, and the fitness feature $q_{h,w}^{i}$ is obtained sequentially through the multi-scale prototype fusion feature extractor and feature adapter G_θ. The anomaly score is computed by the discriminator, and the anomaly score is evaluated:

$$s_{h,w}^{i} = -D\left(q_{h,w}^{i}\right) \tag{9}$$

The anomaly mapping used for anomaly localization in the reasoning process is defined as:

$$S_{AL}(x_i) := \{s_{h,w}^{i} | (h,w) \in h_0 \times w_0\} \tag{10}$$

The spatial resolution of the input samples is obtained by interpolating the anomaly map $S_{AL}(x_i)$. The smooth boundary is obtained by Gaussian filtering with $\sigma = 4$. The maximum anomaly map score is specified to detect the anomaly detection score in each image:

$$S_{AD}(x_i) := \max_{(h,w) \in h_0 \times w_0} s_{h,w}^{i} \tag{11}$$

IV. Proposed Industrial Product Surface Anomaly Detection and Localization Framework

This section proposes a framework for anomaly detection and localization, summarized in the following three steps:

Prototype Initialization. The pre-trained network on ImageNet is used to obtain the feature mappings of the input image $x_i \in \chi_{train}$ at different scales. The K feature mappings randomly selected from $F_j(\chi_{train})$ are updated using k-means clustering. After clustering, the prototype $P_j \in \mathbb{R}^{K \times c^j \times h^j \times w^j}$ at each scale is kept constant for subsequent model training.

Anomaly detection. The samples to be analyzed are input into the trained MPFNet. The network computes the input features and prototype features, facilitating information exchange between different scales through the multi-scale fusion block. Feature adapter then generate target-oriented adapted features. The discriminator is trained using these adapted features along with anomalous features created by adding Gaussian noise to the adapted features. Finally, the anomaly score $s_{h,w}^{i}$ is calculated by the discriminator.

Anomaly localization. The anomaly mapping $S_{AL}(x_i)$ is calculated by the obtained anomaly score $s_{h,w}^{i}$. The spatial resolution of the input samples is obtained by interpolating the anomaly mapping $S_{AL}(x_i)$. The smooth boundary is obtained by Gaussian filtering with $\sigma = 4$. The most responsive point exists in anomalous regions of any size, and the highest scores recorded in the anomaly maps are used as anomaly detection scores for each image, using these anomaly scores we calculate the F1 and thus obtain the desired thresholds for the segmentation results. The pseudo-code for the training process is described in detail in Algorithm 1.

Algorithm 1 MPFNet training pseudo-code

```
# F: Features obtained from pre-trained network
# P: Prototype feature            # E: Feature Extractor
# G: Feature Adapter               # D: Discriminator
# N: i.i.d Gaussian noise         # K: Number of clusters
P = prototype_init(F)
pretrain_init(E)
random_init(G,D)
for x in data_loader:
    o = E(x,P)  # normal features
    q = G(o)  # adapted features
    q_ = q + random(N)  # anomalous features
    loss = loss_func(D(q),D(q_)).mean()
    loss.backward()
    E = E.detach()  # stop gradient
     update(G, D)  # Adam
def prototype_init(F):   # create prototype feature maps
     kmeans = KMeans(n_clusters=K,random_state=0)
     kmeans.fit(F)
     F = kmeans.cluster_centers_
     return F
def loss_func(s,s_):   # loss function
    th_ = -th = 0.5
     return max(0,th-s) + max(0,th_+s_)
```

TABLE I
MPFNET ANOMALY DETECTION AND LOCALIZATION RESULTS ON MVTEC
Remarks: Image-AUROC% on the left (Detection), Pixel-AUROC% on the right (Localization).

Category	Reconstruction-based		Synthesizing-based		Embedding-based			Ours
Model	AE-SSIM	SSPCAB	CutPaste	DRAEM	DRA	CFLOW	PaDiM	MPFnet
Carpet	87/64.7	93.1/92.6	93.9/98.3	97.0/95.5	92.5/98.2	97.6/**99.2**	**99.8**/99.1	98.6/98.3
Grid	94/84.9	99.7/99.5	**100**/97.5	99.9/**99.7**	98.6/87.3	98.1/98.9	96.7/97.3	99.6/98.1
Leather	78/56.1	98.7/96.3	**100**/99.5	**100**/98.5	98.9/93.8	99.9/**99.7**	**100**/99.2	**100**/99.2
Tile	59/17.5	**100**/**99.4**	94.6/90.5	99.6/99.2	**100**/92.3	97.1/96.2	98.1/94.1	**100**/96.6
Wood	73/60.3	98.4/**96.5**	99.1/95.5	99.1/96.4	99.1/84.6	98.7/86.0	99.2/94.9	**99.7**/94.4
Avg.Text.	78/56.7	98.0/96.9	97.5/96.3	99.1/**97.9**	97.8/91.2	98.3/96.0	98.8/96.9	**99.6**/97.3
Bottle	93/83.4	95.6/**99.2**	98.2/97.6	99.2/99.1	**100**/91.3	99.9/97.2	99.1/98.3	**100**/98.6
Cable	82/47.8	92.7/95.1	81.2/90.0	91.8/94.7	94.2/86.6	97.6/97.8	97.1/96.7	**99.0**/**98.5**
Capsule	94/86.0	96.9/90.2	98.2/97.4	**98.5**/94.3	95.1/89.3	97.0/**99.1**	87.5/98.5	**98.5**/99.0
Hazelhut	97/91.6	**100**/**99.7**	98.3/97.3	**100**/**99.7**	**100**/89.6	**100**/98.8	99.4/98.2	**100**/98.4
Metal nut	89/60.3	**100**/99.4	99.9/93.1	98.7/**99.5**	99.1/79.5	98.5/98.6	96.2/97.2	**100**/98.8
Pill	91/83.0	97.4/97.2	94.9/95.7	**98.9**/97.6	88.3/84.5	96.2/**98.9**	90.1/95.7	98.6/98.4
Screw	96/88.7	97.8/**99.0**	88.7/96.7	93.9/97.6	**99.5**/63.2	93.1/98.9	97.5/98.5	98.1/98.6
Toothbrush	92/78.4	97.9/97.3	99.4/98.1	**100**/98.1	87.5/75.5	98.8/**99.0**	**100**/98.8	98.5/98.2
Transistor	90/72.5	90.2/86.9	96.0/93.0	93.1/90.9	88.3/79.1	92.9/98.2	94.4/97.5	**100**/**98.8**
Zipper	88/66.5	**100**/98.4	99.9/**99.3**	99.9/98.8	99.7/96.9	97.1/99.1	98.6/98.5	99.3/98.9
Avg.Obj.	91/75.8	96.9/96.2	95.5/95.8	97.4/97.0	95.1/83.6	97.1/**98.6**	96.0/97.8	**99.2**/**98.6**
Avg.All.	87/69.4	97.3/96.4	96.1/96.0	98.0/97.3	96.0/86.1	97.5/97.7	96.9/97.5	**99.3**/**98.2**

V. CASE STUDY

1. Experimental Details

A. Datasets

Our experiments are mainly conducted using the MVTec AD dataset [27], which was proposed by MVTec in 2019 for detecting and identifying defects and anomalies in industrial manufacturing processes. The dataset comprises a collection of industrial product images that include 5 distinct texture types and 10 varied object types. Each product also suffers many different types of defects, as shown in Fig. 1. The training set has 3629 images, all of which are normal images. The image resizing and center cropping in the experiments are 256×256.

The MVTec LOCO AD dataset [30] consists of structural anomalies and logical anomalies from five different categories containing a total of 3644 images, namely juice bottle, pushpins, screw bag, splicing connectors, and breakfast box. Structural anomalies are manifested in the form of scratches, dents, and contamination. Logical anomalies violate potential constraints.

B. Evaluation Metrics

This paper evaluates the performance of the model by measuring the area under the receiver operating characteristic curve at both the image level, known as Image-AUROC, and the pixel level, referred to as Pixel-AUROC. Image-AUROC is calculated using the anomaly mapping S_{AD} (Eq. 11), and Pixel-AUROC is calculated using the anomaly mapping S_{AL} (Eq. 10). FPS (Frames Per Second), the number of frames that the network is able to complete detection in one second, is directly related to the network's ability to process images. The more frames that can be processed in the same amount of time, the faster the network is able to detect them.

C. Implementation Details

We use the pre-trained WideResnet50 on ImageNet to extract Layer 2 and Layer 3 feature maps with scales of 512×32×32 and 1024×16×16, respectively, and the gradient is not calculated in this process. In the prototype initialization process, given the significant differences in the number of normal samples in different datasets, in order to determine an appropriate number of prototypes, we set the number of prototypes according to a specific proportion of the total number of normal samples and set the value of K to 12%. The patch size in multi-scale fusion is $p \times p$, where $p = 3$. Anomalous features are generated with Gaussian noise σ set to 0.018. The feature dimension of the local features obtained by the multi-scale prototype fusion feature extractor is set to 1536, as shown in Fig. 2. th and th^-, in the loss function, are set to 0.5 and -0.5. The feature adapter and the discriminator both use the Adam optimizer, and the learning rate is set to 1.5×10^{-4} and 2×10^{-4} respectively, and weight decay to 1×10^{-5}. Training epochs is set to 200 and the batch size is 4.

2. Anomaly Detection and Localization on MVTec

We compare MPFnet with reconstruction-based methods (AE-SSIM [28] and SSPCAB [15]), synthesis-based methods (CutPaste [17] and DRAEM [19]), and embedding-based methods (DRA [23], CFLOW [24], and PaDiM [20]). Anomaly detection and localization results of MPFnet as shown in Table I, our proposed method obtains the highest score in 9 out of 15 categories. Regardless of texture or object surface anomaly detection, our proposed method achieves the highest Image-AUROC score of 99.6% and 99.2%, respectively. In addition to that, our proposed method achieves the highest Image-AUROC score and Pixel-AUROC score for both Cable class and Transistor class anomaly detection and localization. Finally, our proposed MPFnet achieves the highest overall average Image-AUROC score of 99.3% in anomaly detection. In terms of anomaly localization, the highest overall average Pixel-AUROC score, i.e., 98.2%, was achieved.

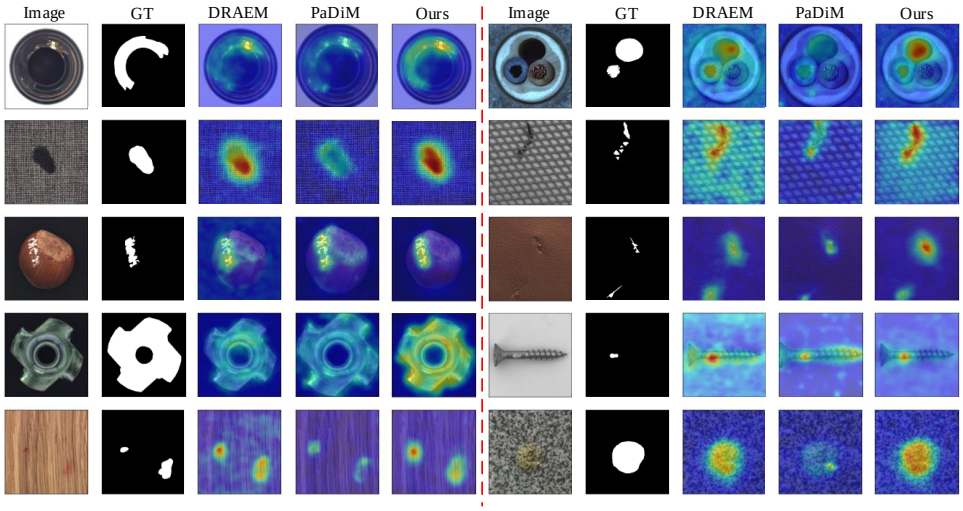


Fig. 3. Comparison of different methods for anomaly localization on MVTec.

We evaluated the anomaly localization performance of MPFnet with the advanced methods DRAEM and PaDiM. The localization visualization results are shown in Fig. 3, which shows that MPFnet can still focus on the anomaly region clearly and locate the anomaly accurately even in the case of subtle defect features. Fig. 4 is an example of MPFnet localization, which shows that the proposed method can achieve good localization results regardless of the defect pattern.

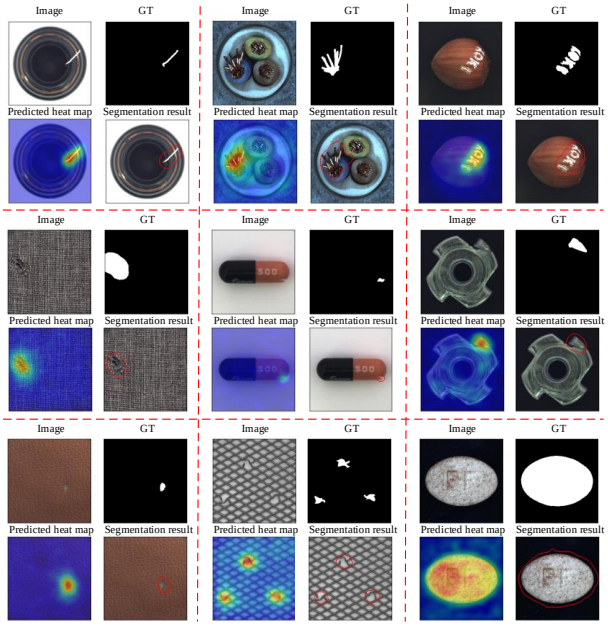


Fig. 4. MPFnet localization example on MVTec.

3. *Anomaly Detection and Localization on MVTec LOCO AD*

To verify the robustness as well as effectiveness of the proposed method, we also conducted experiments on MVTec LOCO AD dataset. We compared MPFnet with VAE [31], DRAEM [19], SimpleNet [9] and FastFlow [32].The logical anomaly detection as well as the structural anomaly detection of MPFnet are shown in Table II. The proposed method is slightly lower than FastFlow in logical anomaly detection but achieves the highest in structural anomaly detection with 84.8%. In addition, our proposed method also achieves the highest overall average anomaly detection at 79.4%.

TABLE II
LOGICAL ANOMALY DETECTION AND STRUCTURAL ANOMALY DETECTION RESULTS

Model	Logical Detection AUROC%	Structural Detection AUROC%	Avg. Detection AUROC%
VAE	53.8	54.8	54.3
DRAEM	72.8	74.4	73.6
SimpleNet	71.5	83.7	77.6
FastFlow	**75.5**	82.9	79.2
Ours	73.9	**84.8**	**79.4**

Fig. 5 shows an example of MPFnet localization on the MVTec LOCO AD dataset, where the proposed method achieves good localization results both in terms of logical anomaly detection and structural anomaly detection.

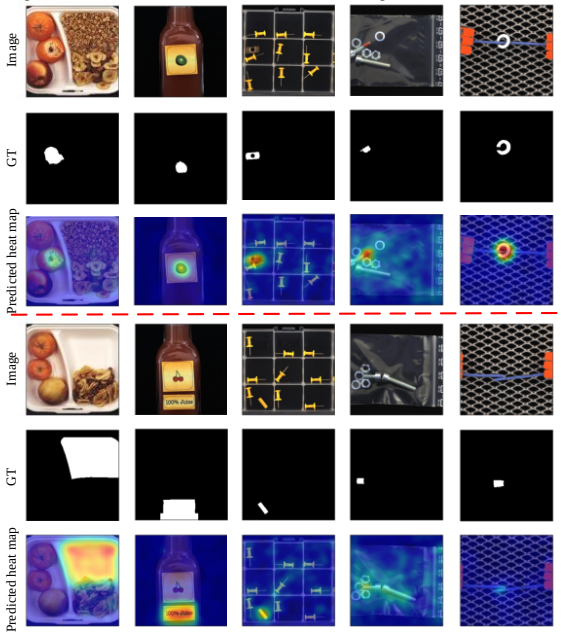


Fig. 5. MPFnet localization examples on MVTec LOCO AD.

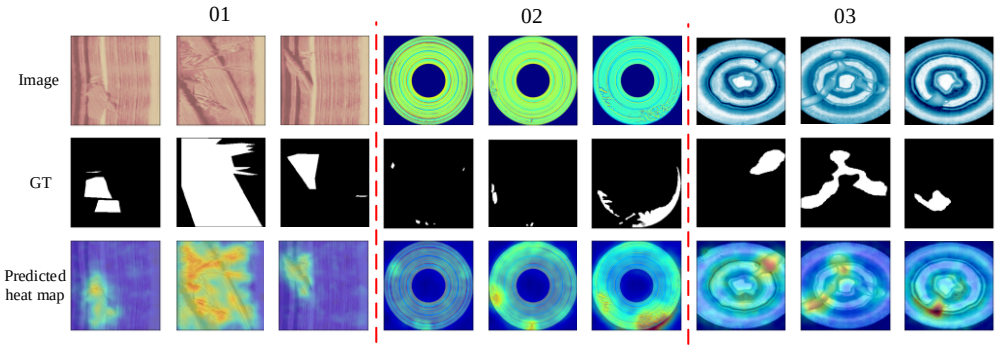


Fig. 6. MPFnet localization examples on BTAD.

4. Anomaly Detection and Localization on BTAD

In real-world industrial application scenarios, it is crucial to validate the proposed method in real-world industrial environments. Therefore, we conducted experiments using the BTAD dataset [33]. beanTech Anomaly Detection (BTAD) dataset is a real-world industrial anomaly detection dataset that contains a total of 2,830 images of real industrial products from three industrial products. We compared MPFnet with SSPCAB [15], DRAEM [19], CFLOW [24] and PatchCore [21]. The Image-AUROC% and Pixel-AUROC% experimental results are shown in Table III. In comparison with other methods, in class 01, the proposed method achieves the highest score of 100% in anomaly detection. In class 02, the proposed method achieves the highest score in both anomaly detection and localization scores, 91.2% and 95.4%, respectively. Finally, in the overall average anomaly detection and localization, the proposed method also achieves the optimal scores of 96.8% and 96.9%, which are higher than the other comparative methods. The BTAD experiments show that, even in the real industrial application scenarios, our proposed method can achieve excellent anomaly detection and localization capabilities.

TABLE III
MPFNET ANOMALY DETECTION AND LOCALIZATION RESULTS ON BTAD

Model	SSPCAB	DRAEM	CFLOW	PatchCore	Ours
01	96.2/92.4	98.5/91.5	93.4/94.9	96.6/**96.5**	**100**/96.1
02	69.3/65.6	68.6/73.4	79.0/93.9	81.3/94.9	**91.2**/**95.4**
03	99.4/92.4	99.8/96.3	99.1/**99.5**	**99.9**/99.2	99.3/99.2
Avg.All.	88.3/83.5	89.0/87.1	90.5/96.1	92.6/96.9	**96.8**/**96.9**

Fig. 6 shows the MPFnet localization visualization results on the BTAD dataset, and it can be further concluded that our proposed method can still focus on the anomalous regions clearly and locate the anomalies accurately even in real and complex industrial application scenarios.

5. Inference time

Reasoning time is a very important issue in industrial model deployment in anomaly detection and localization. Therefore, we have compared the current state-of-the-art methods DRAEM, PaDiM (Resnet18), PaDiM (WideResnet50) with the proposed method MPFnet for different backbone networks as shown in Fig. 7. Our proposed method achieves high Image-AUROC scores along with the highest FPS and the shortest inference time. The experiments in this paper were built in PyTorch (v1.12.1, Python3.10) environment and performed on NVIDIA GeForce RTX 1660 SUPER. In this inference time experiment, all methods batchsize is adjusted to 2.

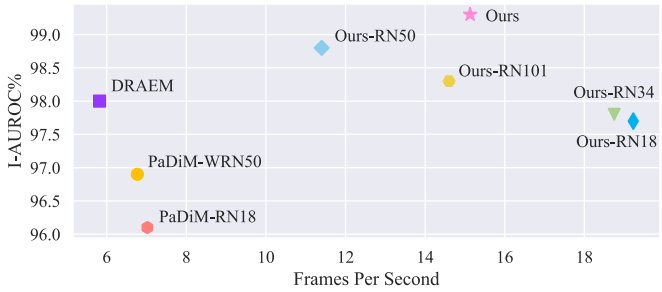


Fig. 7. Inference speed (FPS).

6. Ablation Study

A. Scale of noise

In this paper, we propose to create anomaly samples by introducing Gaussian noise into the adjusted features, while controlling the Euclidean distance of these synthetic anomalies from the normal samples by the scale of the noise. When the noise scale is too small, the discriminator cannot effectively separate the positive anomalous features and the training process is unstable. When the noise scale is too large, it will lead to a loose decision boundary, which in turn leads to high false negatives. As shown in Table IV, MPFnet achieves the highest I-AUROC% and P-AUROC% at a noise scale of 0.018.

TABLE IV
IMPACT OF NOISE SCALE ON THE PROPOSED MODEL PERFORMANCE

Scale of σ	0.010	0.015	0.018	0.025	0.030	0.035	0.050
I-AUROC%	98.8	99.1	**99.3**	98.7	98.6	97.9	96.7
P-AUROC%	97.9	98.1	**98.2**	97.8	97.7	97.4	96.6

B. Neighborhood size

We investigated the effect of the size of the neighborhood p in Eq. 2 on model anomaly detection and localization, as shown in Fig. 8. Too large or too small will reduce the overall model performance of the proposed method, and MPFnet achieves the highest I-AUROC% and P-AUROC% at $p = 3$.

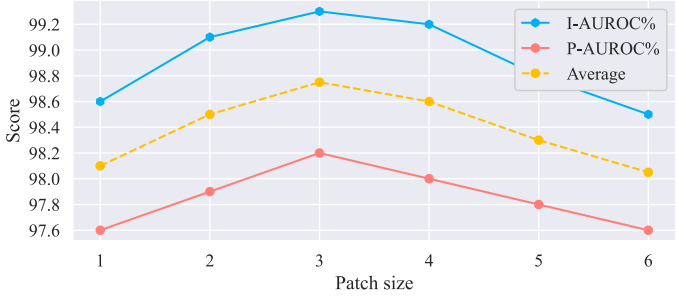


Fig. 8. Performance with varied patch sizes.

C. Number of feature extraction layers

We explored the impact of acquiring different layers of features from the backbone network on the overall model performance, as shown in Table V. It can be found that extracting both Layer 2 and Layer 3 layer features achieved high AUROC scores. In the Layer 2 + Layer 3 model, the proposed method I-AUROC% as well as P-AUROC% reaches the optimization. In addition, the performance of the proposed method decreases when Layer 1 features are added on top of the proposed method, probably because Layer 2 + Layer 3 has basically retained the useful information, and the addition of Layer 1 features affects the discriminator recognition.

TABLE V
IMPACT OF EXTRACTING DIFFERENT LEVELS OF FEATURES ON THE MODEL.

Layer 1	Layer 2	Layer 3	I-AUROC%	P-AUROC%
+			94.2	94.3
	+		96.9	96.6
		+	97.4	97.0
+	+		97.5	96.2
	+	+	**99.3**	**98.2**
+	+	+	99.2	98.0

D. The impact of prototype ratio

We investigated how the ratio of prototypes to the total number of normal samples affects the overall model performance, as detailed in Table VI. 100% implies that the feature mapping of each normal sample is considered as a prototype. A small number of prototypes will result in insufficient differentiation between prototypes, leading to mediocre model performance. When the prototype proportion is set to about 12%, the model performance is basically saturated. Increasing the prototype ratio thereafter results in essentially no improvement in model performance. It is worth noting that the more the number of prototypes the longer the reasoning time, using fewer prototypes can speed up the reasoning, when the prototype ratio is set to 12%, the model performance is basically optimized.

TABLE VI
EFFECT OF PROTOTYPE RATIO ON MODEL PERFORMANCE.

Ratio	5%	10%	12%	20%	50%	100%
I-AUROC%	98.5	99.1	**99.3**	98.8	98.4	98.6
P-AUROC%	97.5	98.1	**98.2**	97.9	97.8	97.8

E. Dependency on backbone

We investigated the effect of different backbone networks on MPFnet, as shown in Fig. 9. The overall model performance of the proposed method is highest when the chosen backbone is WideResnet50.

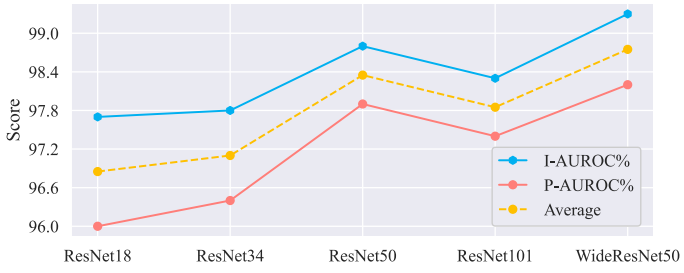


Fig. 9. Performance under different backbones.

F. Inference time

In addition to this, we also investigated the impact of using Resnet18 [29], Resnet34, Resnet50, Resnet101, and WideResnet50 backbone networks in terms of inference time, as shown in Fig. 7. When using the Resnet18 backbone network, the FPS reaches a maximum of 19.23 with an I-AUROC% of 97.7%. When using the WideResnet50 backbone network, the FPS decreased to 15.13, but the I-AUROC% reached a maximum of 99.3%.

VI. CONCLUSIONS

In this paper, we propose a multi-scale prototype fusion network for industrial product surface anomaly detection and localization. First, we propose a multi-scale prototype to represent normal patterns, and construct normal patterns using intermediate feature map prototypes at different scales to provide accurate and representative normal patterns while preserving spatial information. On this basis, the multi-scale fusion block is added to exchange information between different scales, so that the model pays more detailed attention to the prototype features and normal features. Second, feature adapter are constructed to generate target-oriented features and reduce domain bias. The noise scale is calibrated to obtain a tightly bounded normal feature space, and anomalous features are constructed by adding noise to normal features in the feature space. Finally, the anomaly detection process is simplified to be computationally efficient by training a simple discriminator. After extensive experiments on the challenging MVTec AD and MVTec LOCO AD dataset, MPFnet outperforms other state-of-the-art comparative methods in terms of image-level and pixel-level accuracy, as well as shorter inference time. MPFnet can focus on anomalous regions and locate the anomalies accurately even in the case where the defect features are subtle and the defect modes are unknown, and achieves decent detection and localization results.

REFERENCES

[1] D. Luo et al., "Survey on industrial defect detection with deep learning," *Sci. Sin.-Inf.*, vol. 52, no. 6, p. 1002, Jun. 2022.

[2] W. Xie, W. Liu, Y. Xiao, Y. Dai, J. Dong, and R. Liu, "MRSA-Net: Anomaly Detection of Metro Stations Based on Residual Attention Network With Path Aggregation and Efficient Representation," *IEEE Sensors J.*, vol. 23, no. 22, pp. 28340–28354, Nov. 2023.

[3] L. Fan, H. Hu, X. Zhang, H. Wang, and C. Kang, "Magnetic Anomaly Detection Using One-Dimensional Convolutional Neural Network With Multi-Feature Fusion," *IEEE Sensors J.*, vol. 22, no. 12, pp. 11637–11643, Jun. 2022.

[4] S. Mei, J. Cheng, X. He, H. Hu, and G. Wen, "A Novel Weakly Supervised Ensemble Learning Framework for Automated Pixel-Wise Industry Anomaly Detection," *IEEE Sensors J.*, vol. 22, no. 2, pp. 1560–1570, Jan. 2022.

[5] H. Zhang, Z. Wu, Z. Wang, Z. Chen, and Y.-G. Jiang, "Prototypical Residual Networks for Anomaly Detection and Localization," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Vancouver, BC, Canada: IEEE, Jun. 2023, pp. 16281–16291.

[6] P. Perera, R. Nallapati, and B. Xiang, "OCGAN: One-Class Novelty Detection Using GANs With Constrained Latent Representations," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA: IEEE, Jun. 2019, pp. 2893–2901.

[7] S. Han, X. Hu, H. Huang, M. Jiang, and Y. Zhao, "ADBench: Anomaly Detection Benchmark," *ArXiv*, vol. abs/2206.09426, 2022.

[8] J. Deng, W. Dong, R. Socher, L. -J. Li, Kai Li and Li Fei-Fei, "ImageNet: A large-scale hierarchical image database,"*2009 IEEE Conference on Computer Vision and Pattern Recognition*, Miami, FL, USA, 2009, pp. 248-255.
[9] Z. Liu, Y. Zhou, Y. Xu, and Z. Wang, "SimpleNet: A Simple Network for Image Anomaly Detection and Localization," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Vancouver, BC, Canada: IEEE, Jun. 2023, pp. 20402–20411.
[10] M. Yang, P. Wu, and H. Feng, "MemSeg: A semi-supervised method for image surface defect detection using differences and commonalities," *Engineering Applications of Artificial Intelligence*, vol. 119, p. 105835, Mar. 2023.
[11] Q. Zhou, J. Mei, Q. Zhang, S. Wang, and G. Chen, "Semi-supervised fabric defect detection based on image reconstruction and density estimation," *Textile Research Journal*, vol. 91, pp. 962–972, 2020.
[12] P. Bergmann, S. Lwe, M. Fauser, D. Sattlegger, and C. Steger, "Improving Unsupervised Defect Segmentation by Applying Structural Similarity to Autoencoders," 2018.
[13] S. Akcay, A. Atapour-Abarghouei, and T. P. Breckon, "GANomaly: Semi-supervised Anomaly Detection via Adversarial Training," in *Computer Vision – ACCV 2018*, C. V. Jawahar, H. Li, G. Mori, and K. Schindler, Eds., Cham: Springer International Publishing, 2019, pp. 622–637.
[14] S. Akçay, A. Atapour-Abarghouei and T. P. Breckon, "Skip-GANomaly: Skip Connected and Adversarially Trained Encoder-Decoder Anomaly Detection,"*2019 International Joint Conference on Neural Networks (IJCNN)*, Budapest, Hungary, 2019, pp. 1-8.
[15] N. -C. Ristea *et al.*, "Self-Supervised Predictive Convolutional Attentive Block for Anomaly Detection," *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, New Orleans, LA, USA, 2022, pp. 13566-13576.
[16] V. Zavrtanik, M. Kristan, and D. Skočaj, "Reconstruction by inpainting for visual anomaly detection," *Pattern Recognition*, vol. 112, p. 107706, 2021.
[17] C. -L. Li, K. Sohn, J. Yoon and T. Pfister, "CutPaste: Self-Supervised Learning for Anomaly Detection and Localization,"*2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Nashville, TN, USA, 2021, pp. 9659-9669.
[18] J. Song, K. Kong, Y. I. Park, S. Kim, and S. J. Kang, "Anomaly Segmentation Network Using Self-Supervised Learning," 2022.
[19] V. Zavrtanik, M. Kristan, and D. Skocaj, "DRÆM – A discriminatively trained reconstruction embedding for surface anomaly detection," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Montreal, QC, Canada: IEEE, Oct. 2021, pp. 8310–8319.
[20] T. Defard, A. Setkov, A. Loesch, and R. Audigier, "PaDiM: A Patch Distribution Modeling Framework for Anomaly Detection and Localization," in *Pattern Recognition. ICPR International Workshops and Challenges*, A. Del Bimbo, R. Cucchiara, S. Sclaroff, G. M. Farinella, T. Mei, M. Bertini, H. J. Escalante, and R. Vezzani, Eds., Cham: Springer International Publishing, 2021, pp. 475–489.
[21] K. Roth, L. Pemula, J. Zepeda, B. Scholkopf, T. Brox, and P. Gehler, "Towards Total Recall in Industrial Anomaly Detection," *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14298–14308, 2021.
[22] M. Rudolph, T. Wehrbein, B. Rosenhahn, and B. Wandt, "Fully Convolutional Cross-Scale-Flows for Image-based Defect Detection," *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1829–1838, 2021.
[23] C. Ding, G. Pang and C. Shen, "Catching Both Gray and Black Swans: Open-set Supervised Anomaly Detection,"*2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, New Orleans, LA, USA, 2022, pp. 7378-7388.
[24] D. Gudovskiy, S. Ishizaka and K. Kozuka, "CFLOW-AD: Real-Time Unsupervised Anomaly Detection with Localization via Conditional Normalizing Flows,"*2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, Waikoloa, HI, USA, 2022, pp. 1819-1828.
[25] J. Hartingan and M. K. Wong, "Algorithm AS136: A k-means clustering algorithm.," *Applied statistics*, vol. 28, pp. 100–108, 1979.
[26] J. Gu *et al.*, "Multi-Scale High-Resolution Vision Transformer for Semantic Segmentation,"*2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, New Orleans, LA, USA, 2022, pp. 12084-12093.
[27] P. Bergmann, M. Fauser, D. Sattlegger and C. Steger, "MVTec AD — A Comprehensive Real-World Dataset for Unsupervised Anomaly Detection,"*2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, 2019, pp. 9584-9592.
[28] P. Bergmann, S. Löwe, M. Fauser, D. Sattlegger, and C. Steger, "Improving Unsupervised Defect Segmentation by Applying Structural Similarity to Autoencoders," Jan. 2019, pp. 372–380.
[29] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition,"*2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 2016, pp. 770-778.
[30] 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.
[31] D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes," Dec. 10, 2022, arXiv: arXiv:1312.6114. Accessed: Jul. 17, 2024.
[32] J. Yu et al., "FastFlow: Unsupervised Anomaly Detection and Localization via 2D Normalizing Flows," Nov. 16, 2021, arXiv: arXiv:2111.07677. Accessed: Jul. 17, 2024.
[33] P. Mishra, R. Verk, D. Fornasier, C. Piciarelli, and G. L. Foresti, "VT-ADL: A Vision Transformer Network for Image Anomaly Detection and Localization," in 2021 IEEE 30th International Symposium on Industrial Electronics (ISIE), Jun. 2021.

Haidong Shao (Senior Member) received the B.S. degree in Electrical Engineering and Automation and the Ph.D. degree in Vehicle Operation Engineering from Northwestern Polytechnical University, Xi'an, China, in 2013 and 2018, respectively. He is currently an Associate Professor in the College of Mechanical and Vehicle Engineering at Hunan University, Changsha, China. From 2019 to 2021, he was a Postdoctoral Fellow with the Division of Operation and Maintenance Engineering, Luleå University of Technology, Luleå, Sweden. His current research interests include operation and maintenance, data mining, and industrial internet. He has served as the Associate Editors of IEEE Transactions on Industrial Informatics, IEEE Sensors Journal, Journal of Vibration and Control, and IEEE Access.

Jiangji Peng received the B.S. degree in Mechanical Engineering from Lanzhou University of Technology, Lanzhou, China, in 2022. Currently, he is pursuing the M.S. degree in Mechanical Engineering at the School of Mechanical and Vehicle Engineering, Hunan University, Changsha, China. His research interests include industrial anomaly detection

Minghui Shao received her B.S. degree in Mechanical Engineering from Chongqing Jiaotong University, Chongqing, China, in 2023. She is currently working toward the Master degree in Mechanical Engineering with the College of Mechanical and Vehicle Engineering, Hunan University, Changsha, China. Her research interests include the application of unsupervised machine learning algorithms in industrial equipment anomaly detection.

Bin Liu (Member) received the B.S. degree in Automation from Zhejiang University, Hangzhou, China, in 2013, and the Ph.D. degree in Industrial Engineering from City University of Hong Kong, Hong Kong, China, in 2017. He is currently working as a senior lecturer in the Department of Management Science at University of Strathclyde, Glasgow, UK. His research interests include risk analysis, maintenance modeling, prognostic and health management.