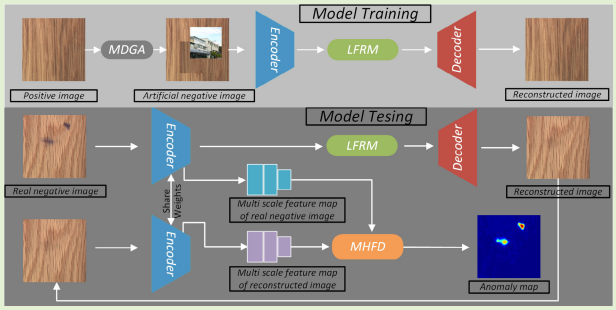
Unsupervised Defect Segmentation via Forgetting-Inputting-Based Feature Fusion and Multiple Hierarchical Feature Difference

Wei Luo, *Student Member, IEEE*, Tongzhi Niu, *Graduate Student Member, IEEE*, Haiming Yao, *Student Member, IEEE*, Lixin Tang, Wenyong Yu, *Member, IEEE*, and Bin Li, *Member, IEEE*

***Abstract*—In the field of surface defect detection, there is a significant imbalance between the number of positive and negative samples, which has led to a growing interest in positive-samples-based anomaly detection methods. Reconstruction-based methods are currently the most commonly used approach, but they often struggle to repair abnormal foregrounds and reconstruct clear backgrounds simultaneously. To address this issue, we propose a new approach called the forgetting-inputting-based feature fusion and multiple hierarchical feature difference network (FIM-Net). The FIM-Net method incorporates a novel latent feature repair module (LFRM), which combines encoding and memory-encoding obtained by memory-augmented module (MAM) via a novel forgetting-inputting-based feature fusion module (FIFFM) to repair abnormal foregrounds while preserving clear backgrounds. Additionally, we introduce a manual defect generation algorithm (MDGA) to simulate realistic and feature-rich anomalies. Finally, we use a multiple hierarchical feature difference (MHFD) for defect segmentation to achieve more accurate defect location. Our extensive comparison experiments demonstrate that the FIM-Net method achieves the state-of-the-art detection accuracy and shows great potential for industrial applications.**



***Index Terms*— Anomaly detection, artificial anomaly images, forgetting-inputting-based feature fusion module (FIFFM), multiple hierarchical feature difference (MHFD), surface defect detection.**

I. Introduction

IN THE industrial field, due to the complexity of the manufacturing process, surface defects are frequently found in various industrial products, including but not limited to fabrics [1] and steel [2], [3]. These defects not only lead to poor user experience but also may cause industrial accidents. For example, surface defects of steel may reduce the contact fatigue strength of the material. Therefore, defects inspection is an important method to achieve quality management. Over the past decades, there have been various surface defect detection methods proposed, which can be broadly categorized into two groups: traditional methods and deep learning-based methods. Traditional methods [4], [5], [6] mainly extract features manually and set thresholds to detect defects. However, the feature extraction ability of the above methods is limited and the robustness is poor. Deep learning is a data-driven method that can automatically extract features by training on large amounts of data. It possesses a strong feature extraction capability and exhibits good generalization.

The majority of deep learning-based methods for surface defect detection are supervised learning approaches [7], [8], [9], [10]. Dong et al. [10] leverage the global contextual information derived from low-resolution feature maps to augment the semantic representation of high-resolution feature maps, thereby enhancing the performance of segmentation tasks. However, it should be noted that these methods require a significant number of defective samples and their corresponding labels, which can be challenging to obtain in industrial

Manuscript received 19 April 2023; accepted 12 May 2023. Date of publication 19 May 2023; date of current version 29 June 2023. This work was supported by the National Natural Science Foundation of China under Grant 51775214. The associate editor coordinating the review of this article and approving it for publication was Prof. Yu-Dong Zhang. *(Wei Luo and Tongzhi Niu contributed equally to this work.) (Corresponding author: Lixin Tang.)*

Wei Luo, Tongzhi Niu, Lixin Tang, Wenyong Yu, and Bin Li are with the State Key Laboratory of Digital Manufacturing Equipment and Technology, School of Mechanical Science and Engineering, Huazhong University of Science and Technology, Wuhan 430074, China (e-mail: u201910709@hust.edu.cn; tzniu@hust.edu.cn; lixintang@hust.edu.cn; ywy@mail.hust.edu.cn; libin999@hust.edu.cn).

Haiming Yao is with the State Key Laboratory of Precision Measurement Technology and Instruments, Department of Precision Instrument, Tsinghua University, Beijing 100084, China (e-mail: yhm22@mails.tsinghua.edu.cn).

Digital Object Identifier 10.1109/JSEN.2023.3276762

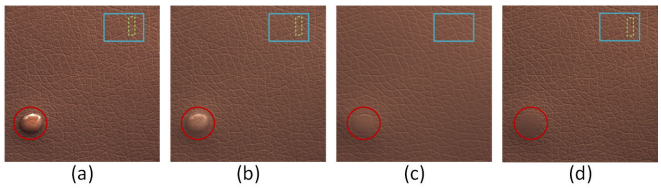


Fig. 1. Reconstruction results of different methods. (a) Original image. (b)–(d) Reconstructed images using AE [14], MemAE [15], and our proposed method FIM-Net, respectively. The red circles and green dotted boxes indicate defects and textures, respectively.

settings. Two main challenges exist in data collection and labeling for defect detection in industry. First, the number of defect-free samples greatly exceeds that of defective samples, and unknown defect types may arise during the production process. Therefore, collecting sufficient data becomes a critical challenge. Second, sample labeling requires skilled engineers to invest considerable time and effort, resulting in high labor costs and time consumption. These factors limit the practical applicability of supervised learning in the industrial field.

Unsupervised learning methods have become increasingly popular due to their ability to train only on unlabeled normal samples. In recent years, unsupervised anomaly detection methods have made significant progress in the field of defect detection. For example, RIAD [11] addresses anomaly detection as a reconstruction-by-inpainting problem and has achieved remarkable results in image anomaly detection. Other methods, such as the deep adversarial anomaly detection method was proposed by Zhang et al. [12] and the normal features distribution model proposed by Cho et al. [13], have also shown excellent results in anomaly detection tasks. These unsupervised anomaly detection methods have shown great potential in defect detection tasks, which is particularly useful in the industrial field where obtaining a large number of defective samples and their corresponding labels can be challenging.

Trained on normal samples, anomaly detection models are expected to generate larger reconstruction or generation errors for anomalies compared to normal samples. Therefore, the two most important capabilities of the model are the ability to reconstruct normal backgrounds and the ability to repair abnormal foregrounds. Common auto-encoder (AE) [14] typically feeds the latent features directly to the decoder, leading to a representation of the latent space that is often under-designed. As a result, these models are capable of the former but not the latter. MemAE [15] proposes an improved AE that repairs the abnormal foregrounds by editing latent codes with a memory-augment module. The proposal of MemAE greatly enhances the repair ability of abnormal foreground but weakens the ability to reconstruct normal background at the same time, as shown in Fig. 1. Specifically, the MemAE method retrieves several relevant memory items for decoding by utilizing the encoding from the encoder as a query, which completely replaces the input features, resulting in a loss of detail. Thus, we introduce a new module called the forgetting-inputting-based feature fusion module (FIFFM) that combines encoding and memory-encoding in a manner that involves forgetting and inputting. This approach helps in repairing abnormal foregrounds while preserving clear backgrounds.

There are two improvements in the FIFFM. At first, inspired by LSTM [16], the way of inputting and forgetting is designed to improve the memory mechanism. We try to erase anomalous foregrounds in coding through a forget gate, and then use the memory-encoded information to inpainting the erased features through an input gate, resulting in a clear reconstruction map. Second, to learn how to forget and input, we proposed a two-stage training strategy. During the initial stage of training with normal samples, the memory content is updated concurrently with the encoder and decoder. In the second stage, the memory contents remain fixed and are no longer updated, and the input and forgetting abilities are learned by repairing artificial anomalies in the artificial abnormal samples.

Recently, artificial anomalies have been widely used to enhance the models. AFEAN [17] generates artificial anomalies by combining defect-free images and random masks. Lv et al. [18] proposed to use redundant features in natural images to simulate defects. Cutpaste [19] cuts a small rectangular area from a normal training image and pastes it back to an image at a random location. But as all as we know, existing artificial anomalies are designed based on human experience and can only simulate limited real anomalies. Therefore, we propose a novel algorithm called manual defect generation algorithm (MDGA) that is capable of generating artificial anomalies that are both realistic and feature-rich. We first assume that the features of natural images are redundant enough to simulate almost all anomalous features. And the blur of normal backgrounds can be regarded as degenerate anomalies. Inspired by Cutpaste, natural images and the blurring images of normal backgrounds as image patches are both pasted at a random location of a large image.

Finally, in anomaly detection, the pixel gap between the original image and the reconstructed image is still used for defect segmentation in anomaly detection, which will cause a lot of noise and lead to the occurrence of false detection. Individual pixel has no semantics, normal and abnormal are context-dependent semantic descriptions. Whether it is filters of traditional methods or convolutional neural networks, contextual relevance is considered to be the key to image processing. Therefore, replacing pixel gaps with feature gaps is a more feasible approach. Since the size of the anomaly is ambiguous, we introduced a new approach for defect segmentation called multiple hierarchical feature difference (MHFD).

There are two advantages in MHFD. First, the residual between the original feature map and the reconstructed feature map guarantees the correlation between pixels, since an element in the feature map corresponds to pixels in one region of the original image. Second, different feature maps have different receptive fields, multiscale feature map residuals are used to obtain multiscale information. In summary, our work makes the following contributions.

1) We introduce a novel FIFFM method, which solves the problem of poor normal background reconstruction in MemAE [15]. A two-stage training strategy is adopted to improve the ability to reconstruct normal background

and the ability to repair abnormal foreground, respectively.

2) We propose a novel MHFD method for defect segmentation, which addresses the noise issues caused by pixel gap and improves defect location accuracy.
3) We also present a novel MDGA algorithm to simulate various defects that can occur in industrial settings.

In this article, we first discuss related works in anomaly detection in Section II. In Section III, we introduce our proposed methods: forgetting-inputting-based feature fusion and MHFD network (FIM-Net). We provide detailed explanations of each method and how they address existing limitations. Section IV presents our experimental results, demonstrating the effectiveness of FIM-Net. Finally, in Section V, we conclude the article.

II. Relate Works

Recently, there has been increasing interest in anomaly detection based on positive samples without labels, as the success of deep learning model training often relies on representative training samples and high-quality annotation. Reconstruction-based methods, including AE [14] and variants of generate adversarial nets (GANs) [20], have become popular for anomaly detection.

AE-based models can learn encoded features in the latent feature domain by training normal samples, and then reconstruct the background images from them. During testing, the output of anomaly samples is expected to be unknown, with a large gap in anomaly regions. To enhance the representation ability, many methods have paid a lot of efforts in the design of network structure and loss function. Mei et al. [21] proposed MSCDAE, which uses a Gaussian pyramid structure to obtain different receptive fields to generate more realistic background images. Yang et al. [22] proposed a fully convolutional AE depend on multi-scale feature clustering for reconstructing image backgrounds. However, these AE-based models only minimize the L1 or L2 loss between the input and reconstructed images, without considering the structural information of the image. As a result, the reconstructed images may be blurred. Therefore, Bergmann et al. [23] proposed a method called AE-SSIM, which incorporates structural similarity into an AE to inspect defects in image backgrounds. However, neural networks have strong generalization ability, which can result in some defects being perfectly reconstructed. And none of the above methods have the ability to deal with abnormal features properly. To solve this problem, MemAE [15] uses a memory-augmented module (MAM) to improve its performance by obtaining reconstruction from selected memory contents of normal data. MemAE can repair the abnormal samples well, amplifying reconstruction errors in abnormal areas. Recently, many improved methods based on memory mechanisms have been proposed, such as TrustMAE [24] and DAAD [25]. However, from the experimental results, MemAE and its variants repair the abnormal foreground well but blur the normal background at the same time.

Recently, GANs [20] have been widely used in various fields, such as style transfer [26] and image generation [27], which shows that GANs have a strong generative ability. Schlegl et al. [28] employed GANS in the field of defect detection. Since the original GANs lacks the mapping from the image domain to the latent feature domain, f-AnoGAN [29], AAE [30], OCGAN [31], and GPND [32] are proposed. Since then, many potential anomaly detection models base on GAN have been proposed. GANomaly [33] proposes an encoder–decoder–encoder framework to reduce the difference between input and reconstructed image in both image and latent feature spaces. Skip-GANomaly [34] improves upon GANomaly by combining skip-connection and GANomaly to reconstruct a more realistic image background with increased details. However, due to the introduction of the encoder-decoder networks structure, the above GANs-based methods are also unable to properly handle abnormal features. Therefore, Yang et al. [17] proposed AFEAN to eliminate the effect of defect reconstruction by editing defect sample features. Niu et al. [35] proposed a memory-augmented adversarial autoencoder for defect detection, which edits the latent features through memory mechanism and ConvLSTM [36]. In general, most of these methods are trained only on positive samples, and localize defects through residuals between original and reconstructed images. However, these methods do not achieve good performance because they exploit the pixel gap between the original image and the reconstructed image to locate defects, resulting in inaccurate defect localization and a large amount of noise.

III. Proposed FIM-Net Methodology

This section provides a detailed introduction to the proposed FIM-Net. First, we provide a brief overview of the overall network architecture. Then, the main modules of FIM-Net are divided into five parts and introduced in detail, including the latent feature repair module (LFRM) comprising MAM and FIFFM, MDGA, the MHFD, the encoder and the decoder, and the two-stage training strategy. Finally, the details of the reconstruction loss and the design of the loss function are discussed.

A. Overall Network Architecture of FIM-Net

The overall structure of the FIM-Net is shown in Fig. 2. The proposed FIM-Net consists of five major components: MDGA, encoder, LFRM, decoder, and MHFD.

The training phase is divided into two stages. In the first stage, our training set contains only positive samples I_p. First, we divide the images into kth patches (patch-size: 64×64) and extract the latent features z_k by encoder. Second, the latent features z_k are re-encoded by LFRM to get memorized latent features $\hat{\hat{z}}_k$. Finally, the $\hat{\hat{z}}_k$ is fed into decoder to get reconstruct images $I_p{}'$. The LFRM, encoder, and decoder are optimized simultaneously. In the second stage, the memory of LFRM is no longer optimized. To make the model better address abnormal features, artificial negative samples I_{an} are generated by MDGA and used as the training set. Through the forgetting and inputting mechanism of LFRM, the abnormal foreground is repaired, the normal background is reconstructed, and finally the images $I_{\text{an}}{}'$ is obtained.

During the testing phase, the negative samples I_n are reconstructed by encoder, LFRM, and decoder to obtain the

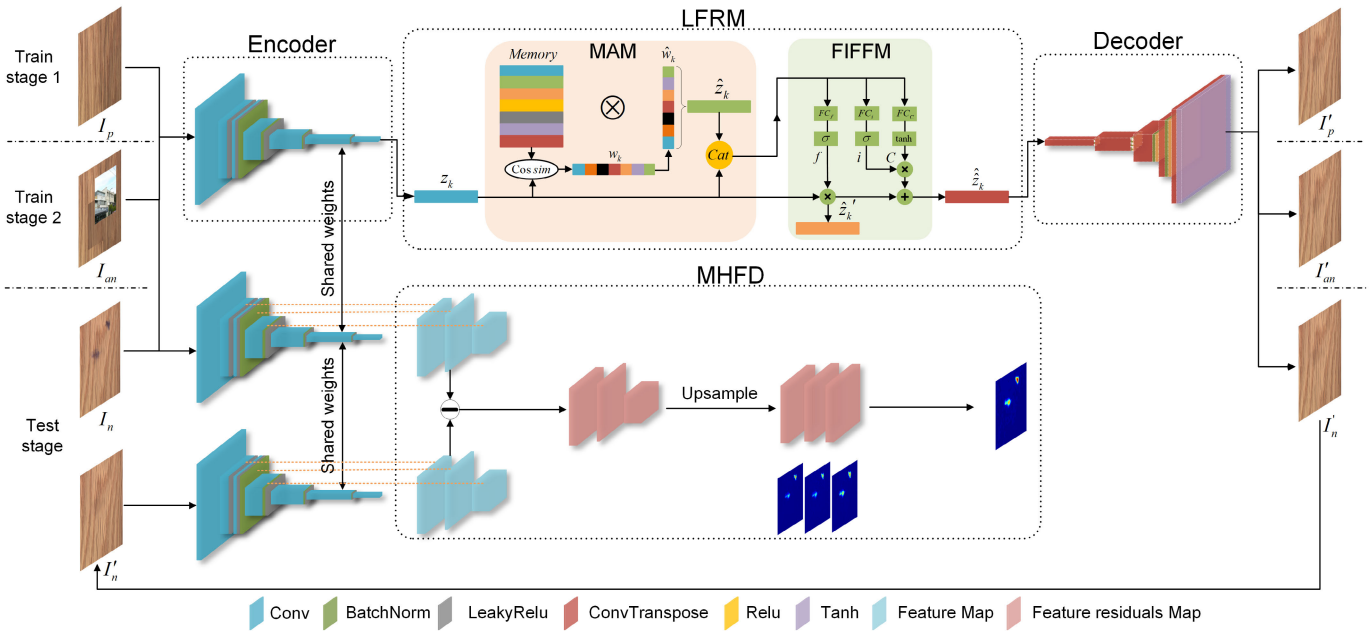


Fig. 2. Overall architecture of the proposed FIM-Net in the train and test stages. FIM-Net consists of a LFRM, a MHFD, an encoder, and a decoder. During training, positive samples and artificial negative samples are propagated forward in two stages. During testing, the abnormal area of negative samples is obtained by MHFD.

images $I_n{}'$. MHFD uses multiscale information of input I_n and output $I_n{}'$ for more accurate anomaly segmentation.

B. Manual Defect Generation Algorithm

Due to the limitations of artificial design, artificial anomaly samples usually can only simulate a few real anomalies. Inspired by Cutpaste [19], we propose MDGA, which involves cutting an image patch and randomly pasting it in a different location of the same image, as shown in Fig. 3. First, based on the feature redundancy of natural images relative to industrial images, we randomly crop a 256 × 256 patch from a natural image I_{na} in the ImageNet dataset [37]. Then, to more closely simulate anomalous features, we also randomly crop the patch from normal samples I_p and resize them to 256 × 256 before blurring. In particular, blurring allows patches to be considered degenerate anomalies, and resizing enables the networks to learn not the ability to deblur, but the ability to repair anomalies. Finally, these two patches are randomly pasted into the normal samples to obtain artificial defect I_{an}, as follows:

$$I_{\text{an}} = \text{paste}\left(I_p, \text{crop}\left(I_{\text{na}}\right), \text{blur}\left(\text{resize}\left(\text{crop}\left(I_p\right)\right)\right)\right). \quad (1)$$

The operations of random crop, resize, and blur are represented by crop(), resize(), and blur(). paste(x_1, x_2, x_3) represents the operation of randomly pasting x_2, x_3 onto x_1.

C. Encoder and Decoder

In the reconstruction task, the images are first mapped to the feature domain using encoder and then mapped to image space by decoder. By setting an information bottleneck, a dimensionality-reduced data representation is obtained in the feature space. Considering the representation ability and computational cost, we design symmetric encoders and decoders, as depicted in Table I.

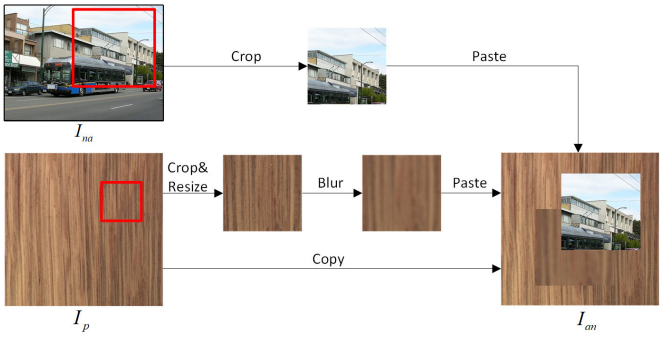


Fig. 3. MDGA. The artificial negative samples are obtained by crop-and-paste of natural images and positive samples.

In the encoder, 4 × 4 convolution kernels with a 2 × 2 strides are used for dimensionality reduction. To increase the receptive field, the second layer of the encoder adopts 3 × 3 convolution kernels with strides of 1 × 1. Symmetric, the decoder uses 4 × 4 deconvolution kernels with a 2 × 2 strides, and the penultimate layer uses 3 × 3 deconvolution kernels with strides of 1 × 1.

It is noteworthy that the proposed framework in this article does not employ skip connections to enhance the reconstruction's level of detail. This is because skip connections would introduce defect features from the encoder to the decoder, resulting in a perfect reconstruction of the defect and leading to false detection.

D. Latent Feature Repair Module

To make the network capable of repairing abnormal foreground and reconstructing normal background at the same time, based on MemAE [15], we propose the LFRM. As shown in Fig. 4, the LFRM includes the MAM and FIFFM. With

TABLE I
FIM-NET ARCHITECTURE

Operation	Kernel	Strides	Padding	Features maps/units	Bn?	Activation function
Encoder						**Input: x=[B,3,64,64]**
Conv	4×4	2×2	1	3→32	×	
Conv	3×3	1×1	1	32→32	√	LeakyReLU
Conv	4×4	2×2	1	32→64	√	LeakyReLU
Conv	4×4	2×2	1	64→128	√	LeakyReLU
Conv	4×4	2×2	1	128→256	√	LeakyReLU
Conv	4×4	1×1	0	256→512	×	
FIFFM				**Input: x_1=[B,512,1,1],x_2=[B,512,1,1], x_{cat}=[x_1,x_2]=[B,1024,1,1]**		
FC				1024→512		Softmax
FC				1024→512		Softmax
FC				1024→512		Tanh
Decoder						**Input: x=[B,512,1,1]**
T.Conv	4×4	1×1	0	512→256	×	
T.Conv	4×4	2×2	1	256→128	√	ReLU
T.Conv	4×4	2×2	1	128→64	√	ReLU
T.Conv	4×4	2×2	1	64→32	√	ReLU
T.Conv	3×3	1×1	1	32→32	√	ReLU
T.Conv	4×4	2×2	1	32→3	×	Tanh
Others:						
				BatchSize(B): 512		
				Optimizer: Adam(α=0.0002, β=0.5)		
				LeaklyReLU: Slope 0.2		

[1] Conv, FC, and T.Conv represent the convolution, full connect, and transposed convolution operations, respectively.

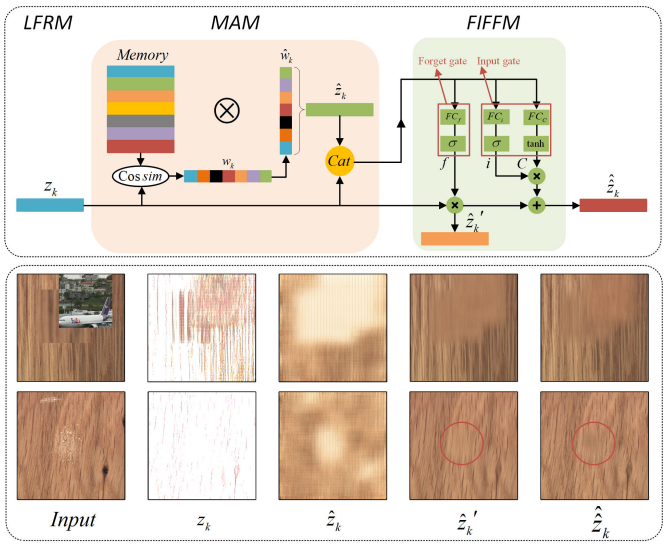


Fig. 4. Diagram of the proposed LFRM. The LFRM includes a MAM and FIFFM. And we show the image obtained by the decoder at each stage of encoding.

MAM, we can get de-anomaly coding $\hat{z_k}$ but lose details. But at the same time, the coding z_k obtained by encoder is rich in detail. Therefore, FIFFM is proposed, which utilizes the forget and input gates of LSTM to better fuse these two complementary feature encodings, resulting in a feature encoding that removes anomalies and contains texture details. Furthermore, to elucidate the effectiveness of the proposed method, as shown in Fig. 4, we feed the encoding obtained at each step into the decoder to obtain the corresponding image.

1) Memory-Augmented Module: In MAM, the encoding from the encoder is used as a query to retrieve the most relevant memory items from the memory bank for decoding to get de-anomaly coding, as shown in Fig. 4.

During training, the memory bank is utilized to store the typical normal patterns, implemented as a matrix $M \in R^{N\times C}$, where N represents the quantity of stored memory items, and C represents the fixed dimension. Specifically, we assign the value of C to be the same as the dimension of the latent feature vector $z_k \in R^C$, which denotes latent feature of the kth patch in the input image. And $m^i \in R^C (i \in \{1, 2, \ldots, N\})$ is used to refer to the ith memory item of memory bank M. We define the memory bank as a content addressable memory [38], [39] with a specific addressing scheme.

The memory coding $\hat{z_k}$ is addressed by the attention weight $w_k \in R^N$, and $w_k^i (i \in \{1, 2, \ldots, N\})$ is obtained by computing the normalized cosine similarity between kth input latent feature z_k and ith memory item m^i in the memory bank, as follows:

$$w_k^i = \frac{\exp\left(\langle z_k, m^i \rangle\right)}{\sum_{j=1}^{N} \exp\left(\langle z_k, m^j \rangle\right)} \tag{2}$$

where $\langle \cdot, \cdot \rangle$ represents the cosine similarity. However, it is still possible to achieve good reconstruction of certain defects through complex linear combinations of many unrelated memory items that contain numerous small elements. Therefore, we employ shrinkage operation. Specifically, items with attention weights less than $1/N$ are removed and re-normalized

$$\hat{w}_k^i = \frac{\max\left(w_k^i - \frac{1}{N}, 0\right) \cdot w_k^i}{|w_k^i - \frac{1}{N}| + \varepsilon} \tag{3}$$

where $\max(\cdot, 0)$ is ReLU activation function, and ε is a very small positive scalar. After the shrinkage operation, we re-normlizaed $\hat{w}_k$ by letting $\hat{w}_k^i = (\hat{w}_k^i / \|\hat{w}_k\|)$. Then memory coding $\hat{z_k}$ is computed as follows:

$$\hat{z}_k = \hat{w}_k M = \sum_{i=1}^{N} \hat{w}_k^i m^i. \tag{4}$$

As suggested in [15], sparse loss function is leveraged to further improve the sparsity of the attention weight $\hat{w}$

$$L_s = \sum_{i=1}^{N} -\hat{w}_k^i \cdot \log\left(\hat{w}_k^i\right). \tag{5}$$

This sparse loss function in (5) and the shrinkage operation in (3) jointly improve the sparsity of the attention weights.

2) Forgetting-Inputting-Based Feature Fusion Module: The coding z_k obtained by the encoder and the de-anomaly coding $\hat{z_k}$ addressed by MAM are complementary in feature domain, where the former is rich in texture but with defective information and the latter is de-anomaly but less rich in texture. Therefore, we concatenate z_k and $\hat{z_k}$ to get $[z_k, \hat{z_k}]$. The process of fusing coding z_k and memory coding $\hat{z_k}$ is represented in (6)–(10). To begin with, as presented in (6)–(7), the forget gate leverages a fully connected layer and a softmax activation function to perform attention operations on the concatenated features, resulting in the forget weights f, which are then multiplied by the original encoding z_k to obtain the feature encoding $\hat{z}_k'$ with removed anomalies. However, $\hat{z}_k'$

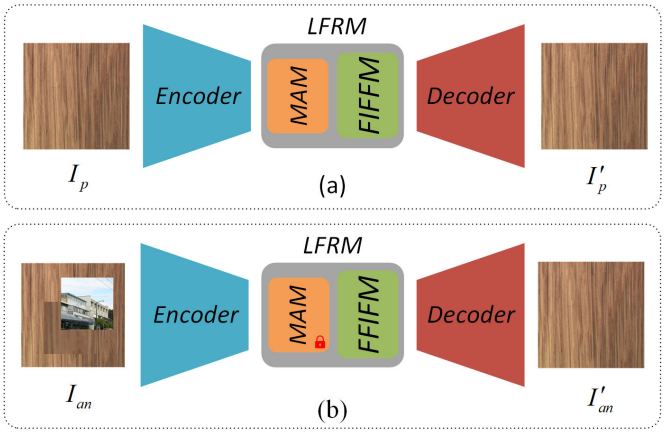


Fig. 5. Illustration of the two-stage training strategy of FIM-Net. (a) Training stage 1, trained with positive samples, the encoder, decoder, MAM, and FIFFM are updated together. (b) Training stage 2, trained with artificial negative samples, MAM is no longer updated.

may lack texture details. Therefore, as depicted in (8)–(10), the input gate employs two fully connected layers, a softmax activation layer, and a tanh activation layer to obtain the input weights i and input texture information C, respectively. These are utilized to augment the texture details in $\hat{z}'_k$, thereby obtaining the feature encoding $\hat{\hat{z}}_k$ with removed anomalies and rich texture information.

Forget gate

$$f = \sigma\left(W_f \cdot \left[z_k, \hat{z}_k\right] + b_f\right) \tag{6}$$

$$\hat{z}'_k = f \cdot z_k. \tag{7}$$

Input gate

$$i = \sigma\left(W_i \cdot [z_k, z] + b_c\right) \tag{8}$$

$$C = \tanh\left(W_c \cdot [z_k, z] + b_c\right) \tag{9}$$

$$\hat{\hat{z}}_k = \hat{z}'_k + i \cdot C \tag{10}$$

where $\sigma(\cdot)$ denotes softmax function, $\tanh(\cdot)$ is tanh activation function, the fully connected layer weights are denoted as W_f, W_i, and W_c, while the biases are b_f, b_i, and b_c.

As depicted in Fig. 4, the reconstructed image by the coding $\hat{z}'_k$ through forget gate is less defective. On the other hand, the reconstructed image using the coding $\hat{\hat{z}}_k$ through the input gate presents a richer texture.

E. Two-Stage Training Strategy

As discussed above, it is difficult to make the model have the ability to reconstruct normal background and repair abnormal foreground at the same time in one stage of training. Therefore, we propose a two-stage training strategy, as shown in Fig. 5. In the first procedure, the model is optimized by training positive samples I_p so that the MAM record prototypical patterns of normal features. In the second procedure, the memory bank in MAM is fixed, and the encoder, decoder, and FIFFM are optimized by training artificial negative samples I_{an} so that the FIFFM learns how to forget defective information and input normal textured information. As shown in Fig. 5(a), in training stage 1, considering that the model optimized with

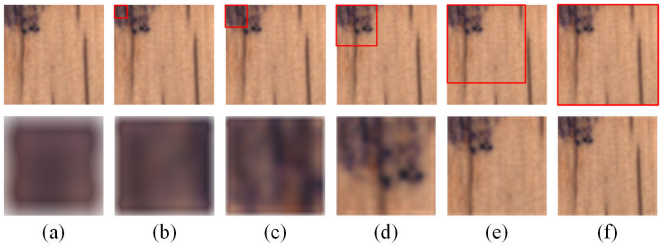


Fig. 6. Different receptive fields of different scale. The red boxes indicate the region of receptive field. (a)–(f) Represent the receptive fields of the feature maps in the encoder from the first to the sixth layer.

L2-norm will produce blurry reconstructed images, L1-norm is used as the criterion for measuring distance

$$L_{\text{rec1}} = \mathop{\mathbb{E}}_{I_p \sim P_{I_p}} \left[||I_p - I_p{}'||_1\right] \tag{11}$$

where $||\cdot||_1$ denotes the L1-norm. To facilitate the sparsity of the attention weights, as described in Section III-D.1, FIM-Net is trained with the sparsity loss L_s. Therefore, the overall model is optimized under the joint loss function

$$L_1 = w_1 L_{\text{rec1}} + w_2 L_s \tag{12}$$

where w_1, w_2 are the weights that control the relative importance of two terms. In this article, we recommend setting $w_1 = 50$ and $w_2 = 0.01$.

As shown in Fig. 5(b), in training stage 2, same as training stage 1, L1-norm is used as the criterion for measuring distance

$$L_{\text{rec2}} = \mathop{\mathbb{E}}_{I_{\text{an}} \sim P_{I_{\text{an}}}} \left[||I_{\text{an}} - I_{\text{an}}{}'||_1\right]. \tag{13}$$

MAM is fixed, encoder, FIFFM, and decoder are optimized under the following joint loss function:

$$L_2 = w_1 L_{\text{rec2}} + w_2 L_s. \tag{14}$$

The values of w_1, w_2 are same as in the training stage 1.

F. Multiple Hierarchical Feature Difference

In fact, anomalies are composite representations of multiple pixels with contextual relationships and have significant multiscale properties. Therefore, to obtain accurate and noise-free anomaly segmentation maps, we propose MHFD. At first, we utilize the differences of feature maps instead of differences of images, which guarantees the correlation between pixels. An element in the feature map corresponds to a region in the original image, which can be thought of as an anomaly score for that region, as shown in Fig. 6. However, as the receptive field expands, an element in the feature map corresponds to a larger area of the image. When there are abnormal areas and normal areas exist at the same time, there will be errors whether the element is represented as normal or abnormal. To address this trade-off, we utilize MHFD to obtain multiscale information.

In many existing methods [10], [24], VGG is trained as a feature extraction network on large-scale datasets, such as ImageNet [37]. These networks generalize well, but do not work for specific data. In this article, the encoder of FIM-Net can be regarded as a feature extractor learned from self-supervised

TABLE II
DETAILED INFORMATION ABOUT DATASETS

Texture Surface Dataset		MV-TAD					DAGM			
		Carpet	Grid	Leather	Tile	Wood	Wallpaper	Bcement	MAGtile	WHcement
Training	W defects number	0	0	0	0	0	0	0	0	0
	W/O defects number	280	264	245	230	247	275	275	275	275
Testing	W defects number	88	57	72	84	60	300	150	150	300
	W/O defects number	21	21	32	33	19	0	0	0	0

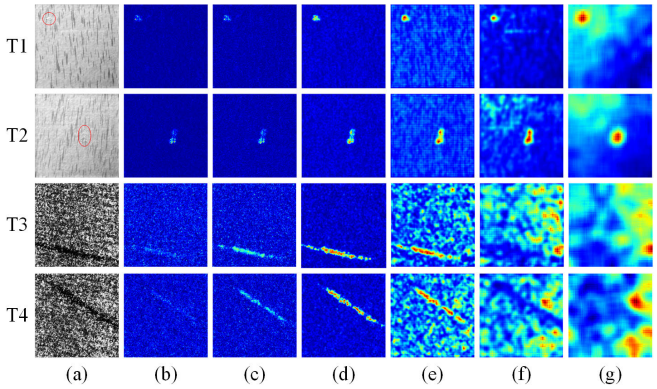


Fig. 7. Feature differences by the proposed MHFD. (a) Defective images. (b)–(g) Error maps obtained by scale 1, scale 2, scale 3, scale 4, scale 5, and scale 6.

learning on a specific dataset. As shown in Fig. 2, in MHFD, the images I_n and the corresponding reconstructions $I_n{}'$ are fed to the encoder, so we can get the multiscale feature maps. However, not all feature maps can clearly express abnormal regions. We perform some experiments and analysis on the selection of feature maps. As illustrated in Table I, the encoder has a total of six convolutional layers. As depicted in Fig. 7, the error maps of scale 4, 5, 6 have a lot of noise. In this article, we select the feature maps of scale 1, 2, 3 to obtain anomaly segmentation results.

Then, the feature maps subtracted and squared, upsampled to the size of the original image, and averaged over the channel dimension. Finally, a weighted average of feature difference at multiple scales is performed. The images is represented by $I_n, I_n{}' \in R^{C\times H\times W}$. Let $\phi_l(I) \in R^{C_l\times H_l\times W_l}$ represents the lth layer feature map. The MHFD$(I_n, I_n{}') \in R^{H\times W}$ can be described as follows:

$$\mathrm{MHFD}\left(I_n, I_n{}'\right) = \sum_{l=1}^{3} \lambda_l \tau\left(\|\phi_l\left(I_n\right) - \phi_l\left(I_n{}'\right)\|_2\right) \quad (15)$$

where $\tau(\cdot)$ is a bilinear upsampling function that resizes the feature tensors to $H \times W$, $\|\cdot\|_2$ denotes the L_2 distance, and λ_l denotes the corresponding weight. In terms of weight setting, we devised a relatively scientific approach. Different feature maps correspond to different receptive fields, and we postulate that larger receptive fields contribute to better defect discrimination. To this end, we denote the receptive field size of the lth layer feature map as RF_l. Accordingly, the corresponding weight λ_l can be expressed as

$$\lambda_l = \frac{RF_l}{\sum_{j=1}^{3} RF_j}. \quad (16)$$

IV. EXPERIMENTATION

A. Set Up

In this section, the effectiveness of the proposed FIM-Net method is evaluated through several sets of experiments.

1) The overall inspection performance of FIM-Net is compared with nine state-of-the-art methods on five benchmark textured surfaces in the MV-TAD dataset [40].
2) To further verify the generalizability of the model, a comparative experiment on four challenging textured surfaces in the DAGM dataset [41] is conducted.
3) The influences of each component in the FIM-Net are explored in ablation experiments.
4) The inference speed of FIM-Net is compared with other outstanding methods.
5) The FIM-Net is evaluated on an industrial dataset to validate its industrial potential.

In these experiments, a variety of anomaly detection samples are used, including carpet, grid, leather, tile, wood, wallpaper, Bcement, MAGtile, and WHcement. The carpet, grid, leather, tile, and wood textured surfaces are sourced from MV-TAD [40], and the wallpaper, Bcement, MAGtile, and WHcement textured surfaces are sourced from DAGM [41]. The defect samples in the MV-TAD dataset exhibit significant variability in terms of color, shape, and scale, making it a highly challenging dataset. In contrast, the defect samples in the DAGM dataset have highly intricate backgrounds, which pose a significant challenge in distinguishing between defects and complex textures. In our experiments, all images were resized to 512 × 512 pixels. Additional detailed information regarding the datasets is summarized in Table II.

To quantitatively analyze the performance of various methods, we adopt the area under the receiver operating characteristic curve (AuROC) as evaluation criterion, which is insensitive to thresholds and can better evaluate the inspection performance of models.

All the experiments are implemented using Python 3.8.0 and Pytorch 1.9.1 on a computer with an NVIDIA Tesla A100 GPU, which is equipped with 40 Intel(R) Xeon(R) CPU E5-2640 v4 at 2.40 GHz and 40 GB memory.

B. Overall Performance Comparison on MV-TAD Dataset

We compare the inspection performance of the proposed FIM-Net method with several outstanding anomaly detection methods to verify its overall effectiveness, including AE_SSIM [23], AnoGAN [28], f-AnoGAN [29], MS-FCAE [22], MemAE [15], RIAD [11], TrustMAE [24], VAE [42], ACDN [43], and AFEAN [17].

TABLE III
AUROC RESULTS ON FIVE CATEGORIES OF TEXTURED SURFACES IN MV-TAD DATASET

Category	AE-SSIM	AnoGAN	f-AnoGAN	MS-FCAE	MemAE	TrustMAE	RIAD	VAE	ACDN	AFEAN	**FIM-Net**
Carpet	87.00	54.00	66.00	78.20	81.16	**98.53**	96.30	73.50	91.10	90.30	91.25
Grid	94.00	58.00	85.00	88.10	95.56	97.45	98.80	96.10	94.10	92.60	**99.31**
Leather	78.00	64.00	83.00	91.70	92.91	98.05	**99.40**	92.50	98.40	96.10	99.13
Tile	59.00	50.00	72.00	53.20	70.76	82.48	89.10	65.40	93.60	85.70	**98.82**
Wood	73.00	62.00	74.00	81.20	85.44	92.62	85.80	83.80	92.90	92.20	**96.96**
Ave.	78.00	58.00	76.00	78.50	85.17	93.83	93.70	82.26	94.10	91.40	**97.09**

[1] The best AuROC result is in bold, and the second best is underlined.

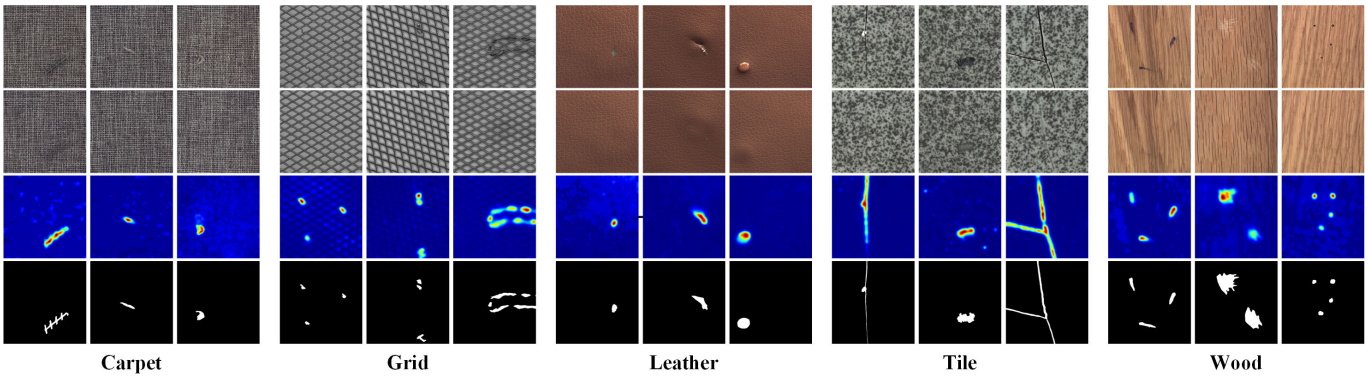


Fig. 8. Inspection results for five types of textures in MV-TAD [40]. From top to bottom are the input defective images, the reconstruction images, the error maps, and the ground truth, respectively.

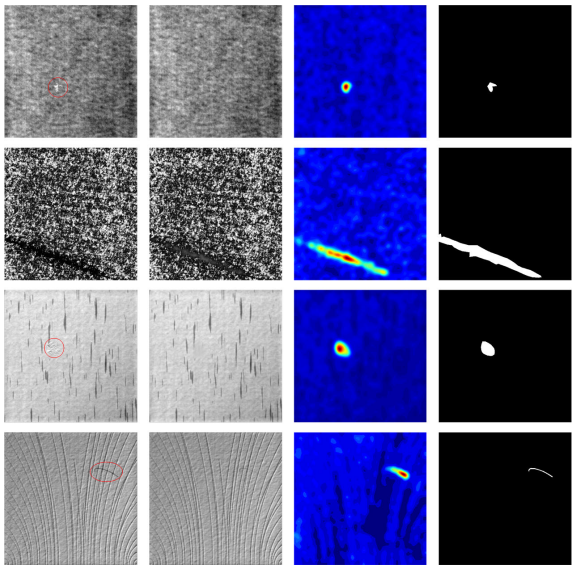


Fig. 9. Inspection results for four types of textures in DAGM [41].

TABLE IV
AUROC RESULTS ON FOUR CATEGORIES OF TEXTURED SURFACES IN DAGM DATASET

Category	MAGtile	Bcement	wallpaper	WHcement	Ave.
AE-SSIM	74.30	79.40	95.60	81.70	82.75
CNN_Dict	69.20	75.80	66.50	67.90	69.85
AnoGAN	78.10	45.50	72.20	72.70	67.13
OCGAN	89.10	74.50	97.10	95.60	89.08
MS-FCAE	95.90	58.80	90.50	**96.40**	85.40
AFEAN	97.40	**94.80**	98.30	96.10	96.65
FIM-Net	**99.97**	90.45	**99.90**	96.35	**96.67**

[1] The best AuROC result is in bold, and the second best is underlined.

The experimental results of quantitative analysis are presented in Table III. FIM-Net achieves a better average result compared to other outstanding methods. Especially for grid, tile, and wood, compared to the second-best result, FIM-Net improves AuROC metric by margins of 0.51%, 5.22%, and 4.06%, respectively. FIM-Net performs well on five different samples, which shows that our method can maintain performance on a variety of samples.

Some inspection results of FIM-Net on five different samples are shown in Fig. 8. The FIM-Net leverages MDGA and LFRM to enable the model can simultaneously repair anomalous foregrounds and reconstruct normal backgrounds, thus obtaining clear reconstructed images. Instead of using pixel difference between input and reconstructed images, FIM-Net adopts MHFD to obtain more accurate defect segmentation maps.

The superior results mentioned above can be attributed to three key factors. First, the artificially generated defects by MGDA are able to fit the real defect features in the feature domain, thus enhancing the model's ability to handle out-of-distribution samples. Second, FIFFM effectively fuses two complementary features, input coding and memory coding, through the forget and input gates of LSTM, enabling the model to accurately reconstruct the texture background while repairing defects. Finally, MHFD utilizes feature differences at multiple scales to obtain a detection map with semantic information. In comparison to pixel differences, MHFD can more accurately localize defects and suppress noise, thereby contributing to the improved performance of the proposed method.

C. Inspection Generalizability Experiment on DAGM Dataset

To further verify the generalizability of FIM-Net, the detection performance of FIM-Net is compared with various

TABLE V
AVERAGE INFERENCE TIME

Method	AE-SSIM	f-Ano GAN	Mem AE	Trust MAE	RIAD	FIM-Net
Times (ms)	2.506	7.284	8.840	12.218	67.405	41.673

TABLE VI
ABLATION ANALYSIS FOR THE FIM-NET ON THE LEATHER DATASET

Module	A	B	C	D	E
Encoder	√	√	√	√	√
LFRM(√)/MAM(*)/Cat(#)	√	*	#	√	√
Decoder	√	√	√	√	√
Two-stage training strategy	√	√	√		√
MHFD	√	√	√	√	
AuROC	99.13	90.80	98.71	98.24	98.42

excellent methods, including AE-SSIM [23], CNN_Dict [44], AnoGAN [28], OCGAN [31], MS-FCAE [22], and AFEAN [17].

The results of our quantitative experiments are displayed in Table IV. FIM-Net outperforms other outstanding methods in terms of AuROC metric, which reveals that the proposed method can maintain excellent performance on different datasets, with good generalization.

Fig. 9 shows some defect inspection examples of FIM-Net. The FIM-Net can inspect and locate defective regions accurately on four types of textured surfaces.

D. Inference Time Comparison

A good balance between inference speed and inspection performance is the essential point in practical industrial defect detection. To demonstrate that FIM-Net method can achieve the balance, the inference time of FIM-Net is compared with that of other outstanding methods, comprising AE-SSIM [23], f-AnoGAN [29], MemAE [15], TrustMAE [24], and RIAD [11]. The inference time is evaluated with 512 × 512 pixels resolution.

The quantitative experimental results regarding inference time are shown in Table V. The inference time of FIM-Net is 41.673 ms, ranked behind AE-SSIM, f-AnoGAN, MemAE, and TrustMAE. However, the inspection performances of these methods are inferior to that of FIM-Net method. RIAD method leverages masked images at different scales to inspect defects, leading to slow inference speed, which limits its practical industrial applications. Accordingly, from the perspective of the balance between inference speed and inspection performance, the FIM-Net method is in line with industrial demands.

E. Ablation Analysis

We conducted a set of ablation experiments to determine the impact of each module in FIM-Net. To ensure a fair comparison, all variants of FIM-Net used the same parameter settings. The results of the experiments are presented in Table VI and Fig. 10.

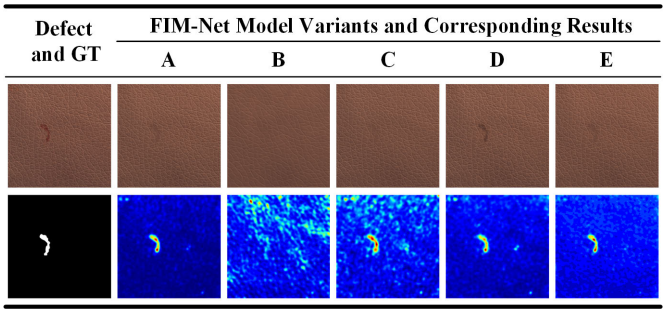


Fig. 10. Examples of images from tests in the ablation study. The first row shows the reconstruct images, and the second row shows the error maps. Each column corresponds to the model variant in Table VI.

1) Influence of LFRM: LFRM is proposed to solve the problem that MemAE [15] enhances the repair ability of abnormal foregrounds but weakens the ability to reconstruct normal backgrounds at the same time. To verify whether our improvement is effective, the impact of replacing LFRM with MAM will be analyzed in detail.

As shown in columns A and B of Fig. 10, after replacing LFRM with MAM, we find that its background is not reconstructed well, resulting in a lot of noise during defect segmentation. Table VI presents the quantitative experimental results. FIM-Net (column A) improves the AuROC by a margin of 8.33% compared to the model (column B) that replaced LFRM with MAM.

To verify that LFRM improves MAM not only due to the increase of parameters, we design a set of comparative experiments between FIM-Net and Cat (replacing forget and input gates with concatenate). As shown in columns A and C of Fig. 10, we find that the error map of LFRM is clearer than that of Cat. Table VI shows the quantitative results. FIM-Net (column A) improves the AuROC by 0.42% compared to the Cat (column C).

This experiment confirms that our LFRM outperforms MAM. LFRM can repair abnormal foreground through the way of forgetting and inputting and can reconstruct normal background well, helping us achieve more accurate anomaly detection.

2) Influence of Two-Stage Training Strategy: The two-stage training strategy aims to let LFRM learn how to forget and input, giving the model the ability to distinguish abnormal foreground from normal background. To confirm its effectiveness, we removed the two-stage training strategy from FIM-Net.

Columns A and D of Fig. 10 show an example of the influence of two-stage training strategy. Without a two-stage training strategy, LFRM cannot learn how to forget and input, resulting in goon reconstruction of defects in the testing phase. The quantitative experimental results are illustrated in Table VI; compared with model without the two-stage training strategy (column D), FIM-Net (column A) improves the AuROC by a margin of 0.89%.

The experiment confirms that our two-stage training strategy is effective. The two-stage training strategy can make the model have the ability to distinguish abnormal foregrounds

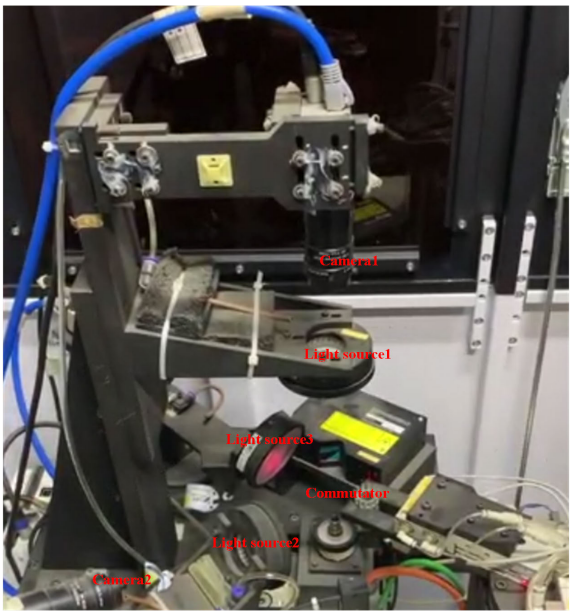


Fig. 11. Automated optical inspection equipment for commutator surface defect detection.

from normal backgrounds, which helps the model to better repair abnormal foregrounds.

3) Influence of MHFD: The purpose of MHFD is to accurately segment defects and suppress noise by leveraging feature difference at multiscales. To verify the effectiveness, we removed the entire MHFD from FIM-Net.

Column A and E of Fig. 10 show an example of the influence of MHFD. Instead of using MHFD, the model (column E) utilizes the pixel difference between input defective and its reconstruction, resulting in inaccurate defect segmentation and a lot of noise. By leveraging MHFD, noise can be effectively suppressed and defect segmentation in more accurate. Table VI presents the quantitative experimental results. FIM-Net (column A) improves the AuROC by 0.71% compared to the model without MHFD (column E).

This experiment confirms that our proposed MHFD is indeed effective. By exploiting MHFD no longer focus on small gap between pixels, but on anomaly scores in regions of different sizes. This helps us achieve more accurate defect segmentation and suppress noise.

F. Industrial Application

To validate the potential of FIM-Net in the industrial field, as shown in Fig. 11, it is implemented in our automated optical inspection equipment to inspect commutator surface defects online, which comprises the camera, light source, commutator, etc. A commutator dataset comprising 443 defect-free images and 66 defective images was collected, of which 336 nominal samples are leveraged for training and 97 nominal samples and 66 anomalous samples are utilized for testing.

Some examples of the defect segmentation results obtained using the FIM-Net method are shown in Fig. 12. The proposed FIM-Net is capable of inspecting and locating various defects accurately, which reveals its potential in the particular application.

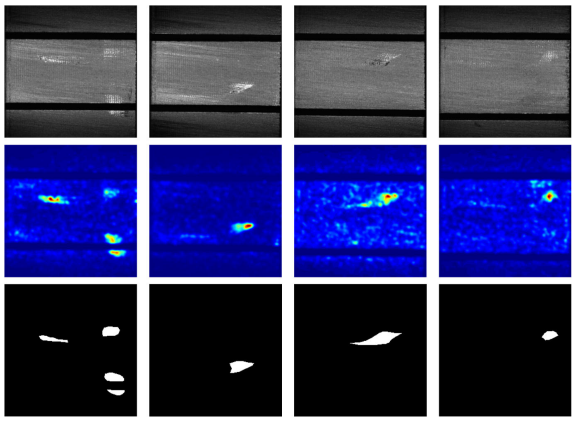


Fig. 12. Examples of commutator surface defect segmentation results of proposed FIM-Net. From top to bottom are the input defective images, the error maps, and the ground truth, respectively.

V. Conclusion

In this article, we proposed an unsupervised learning method FIM-Net for surface defect detection. This method is only trained on positive samples and artificial negative samples without any real negative samples. It is very important for the initial stage of the industrial production lines where negative samples are extremely scarce.

Academically, we divide the key points of reconstruction-based anomaly detection methods into anomaly foreground repair ability and normal background reconstruction ability. A novel LFRM module and a new method of two-stage training strategy are proposed to obtain the two abilities. Furthermore, MHFD is proposed for more accurate and reasonable defect segmentation. We observed that MHFD outperforms the traditional method based on pixel gaps between original image and reconstructed image. Because MHFD uses encoders as feature extractors, it is suitable for all the AE-based or GAN-based methods. Extensive experimental results on several typical anomaly detection datasets show that our method FIM-Net achieves the state-of-the-art detection accuracy.

However, it should be noted that the patch-level reconstruction technique employed in the proposed method may result in a decreased inference speed of the network. Additionally, the approach exhibits limitations in its scope of application, as it is exclusively effective in detecting texture defects and falls short of delivering satisfactory performance in detecting component defects such as screws and transistors. In future work, our focus will be on optimizing the model's inference speed without compromising its performance and investigating alternative approaches to expand the range of applications of our method to include component defect detection.

References

[1] K. Zhang, Y. Yan, P. Li, J. Jing, X. Liu, and Z. Wang, "Fabric defect detection using salience metric for color dissimilarity and positional aggregation," *IEEE Access*, vol. 6, pp. 49170–49181, 2018.

[2] H. Wang, J. Zhang, Y. Tian, H. Chen, H. Sun, and K. Liu, "A simple guidance template-based defect detection method for strip steel surfaces," *IEEE Trans. Ind. Informat.*, vol. 15, no. 5, pp. 2798–2809, May 2019.

[3] X. Wen, J. Shan, Y. He, and K. Song, "Steel surface defect recognition: A survey," *Coatings*, vol. 13, no. 1, p. 17, Dec. 2022.

[4] D. Aiger and H. Talbot, "The phase only transform for unsupervised surface defect detection," in *Proc. Comput. Vis. Pattern Recognit.*, 2010, pp. 295–302.

[5] X. Xie and M. Mirmehdi, "TEXEMS: Texture exemplars for defect detection on random textured surfaces," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 8, pp. 1454–1464, Aug. 2007.

[6] D.-M. Tsai and T.-Y. Huang, "Automated surface inspection for statistical textures," *Image Vis. Comput.*, vol. 21, no. 4, pp. 307–323, Apr. 2003.

[7] L. Xiao, B. Wu, and Y. Hu, "Surface defect detection using image pyramid," *IEEE Sensors J.*, vol. 20, no. 13, pp. 7181–7188, Jul. 2020.

[8] S. Tian et al., "CASDD: Automatic surface defect detection using a complementary adversarial network," *IEEE Sensors J.*, vol. 22, no. 20, pp. 19583–19595, Oct. 2022.

[9] W. Xuan, G. Jian-She, H. Bo-Jie, W. Zong-Shan, D. Hong-Wei, and W. Jie, "A lightweight modified YOLOX network using coordinate attention mechanism for PCB surface defect detection," *IEEE Sensors J.*, vol. 22, no. 21, pp. 20910–20920, Nov. 2022.

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

[11] V. Zavrtanik, M. Kristan, and D. Skočaj, "Reconstruction by inpainting for visual anomaly detection," *Pattern Recognit.*, vol. 112, Apr. 2021, Art. no. 107706.

[12] X. Zhang, J. Mu, X. Zhang, H. Liu, L. Zong, and Y. Li, "Deep anomaly detection with self-supervised learning and adversarial training," *Pattern Recognit.*, vol. 121, Jan. 2022, Art. no. 108234.

[13] M. Cho, T. Kim, W. J. Kim, S. Cho, and S. Lee, "Unsupervised video anomaly detection via normalizing flows with implicit latent features," *Pattern Recognit.*, vol. 129, Sep. 2022, Art. no. 108703.

[14] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, Jul. 2006.

[15] D. Gong et al., "Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1705–1714.

[16] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[17] H. Yang, Q. Zhou, K. Song, and Z. Yin, "An anomaly feature-editing-based adversarial network for texture defect visual inspection," *IEEE Trans. Ind. Informat.*, vol. 17, no. 3, pp. 2220–2230, Mar. 2021.

[18] C. Lv, F. Shen, Z. Zhang, D. Xu, and Y. He, "A novel pixel-wise defect inspection method based on stable background reconstruction," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–13, 2021.

[19] C.-L. Li, K. Sohn, J. Yoon, and T. Pfister, "CutPaste: Self-supervised learning for anomaly detection and localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 9664–9674.

[20] I. Goodfellow et al., "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 1–11.

[21] S. Mei, H. Yang, and Z. Yin, "An unsupervised-learning-based approach for automated defect inspection on textured surfaces," *IEEE Trans. Instrum. Meas.*, vol. 67, no. 6, pp. 1266–1277, Jun. 2018.

[22] H. Yang, Y. Chen, K. Song, and Z. Yin, "Multiscale feature-clustering-based fully convolutional autoencoder for fast accurate visual inspection of texture surface defects," *IEEE Trans. Autom. Sci. Eng.*, vol. 16, no. 3, pp. 1450–1467, Jul. 2019.

[23] P. Bergmann, S. Löwe, M. Fauser, D. Sattlegger, and C. Steger, "Improving unsupervised defect segmentation by applying structural similarity to autoencoders," 2018, *arXiv:1807.02011*.

[24] D. S. Tan, Y. Chen, T. P. Chen, and W. Chen, "TrustMAE: A noise-resilient defect classification framework using memory-augmented auto-encoders with trust regions," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 276–285.

[25] J. Hou, Y. Zhang, Q. Zhong, D. Xie, S. Pu, and H. Zhou, "Divide-and-assemble: Learning block-wise memory for unsupervised anomaly detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 8791–8800.

[26] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4401–4410.

[27] A. Fruhstuck, K. K. Singh, E. Shechtman, N. J. Mitra, P. Wonka, and J. Lu, "InsetGAN for full-body image generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 7723–7732.

[28] T. Schlegl, P. Seeböck, S. M. Waldstein, U. Schmidt-Erfurth, and G. Langs, "Unsupervised anomaly detection with generative adversarial networks to guide marker discovery," *Inf. Process. Med. Imag.*, 2017.

[29] T. Schlegl, P. Seeböck, S. M. Waldstein, G. Langs, and U. Schmidt-Erfurth, "F-AnoGAN: Fast unsupervised anomaly detection with generative adversarial networks," *Med. Image Anal.*, vol. 54, pp. 30–44, May 2019.

[30] A. Makhzani, J. Shlens, N. Jaitly, I. Goodfellow, and B. Frey, "Adversarial autoencoders," 2015, *arXiv:1511.05644*.

[31] P. Perera, R. Nallapati, and B. Xiang, "OCGAN: One-class novelty detection using GANs with constrained latent representations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2898–2906.

[32] S. Pidhorskyi, R. Almohsen, and G. Doretto, "Generative probabilistic novelty detection with adversarial autoencoders," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018, pp. 1–12.

[33] S. Akcay, A. Atapour-Abarghouei, and T. P. Breckon, "GANomaly: Semi-supervised anomaly detection via adversarial training," in *Proc. Asian Conf. Comput. Vis.*, 2018, pp. 622–637.

[34] S. Akçay, A. Atapour-Abarghouei, and T. P. Breckon, "Skip-GANomaly: Skip connected and adversarially trained encoder–decoder anomaly detection," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2019, pp. 1–8.

[35] T. Niu, B. Li, W. Li, Y. Qiu, and S. Niu, "Positive-sample-based surface defect detection using memory-augmented adversarial autoencoders," *IEEE/ASME Trans. Mechatronics*, vol. 27, no. 1, pp. 46–57, Feb. 2022.

[36] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-C. Woo, "Convolutional LSTM network: A machine learning approach for precipitation nowcasting," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, 2015, pp. 1–9.

[37] O. Russakovsky et al., "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.

[38] J. Weston, S. Chopra, and A. Bordes, "Memory networks," in *Proc. 3rd Int. Conf. Learn. Represent. (ICLR)*, 2015, pp. 1–15.

[39] J. Rae et al., "Scaling memory-augmented neural networks with sparse reads and writes," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 29, 2016, pp. 1–9.

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

[41] M. Wieler and T. Hahn. (Jun. 25, 2017). *Weakly Supervised Learning for Industrial Optical Inspection*. [Online]. Available: https://hci.iwr.uniheidelberg.de/node/3616

[42] D. Dehaene, O. Frigo, S. Combrexelle, and P. Eline, "Iterative energy-based projection on a normal data manifold for anomaly localization," in *Proc. Int. Conf. Learn. Represent.*, 2019, pp. 1–17.

[43] K. Song, H. Yang, and Z. Yin, "Anomaly composition and decomposition network for accurate visual inspection of texture defects," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–14, 2022.

[44] P. Napoletano, F. Piccoli, and R. Schettini, "Anomaly detection in nanofibrous materials by CNN-based self-similarity," *Sensors*, vol. 18, no. 2, p. 209, Jan. 2018.

Wei Luo (Student Member, IEEE) is pursuing the B.S. degree with the School of Mechanical Science and Engineering, Huazhong University of Science and Technology, Wuhan, China.

His research interests include deep learning, anomaly detection, and machine vision.

Tongzhi Niu (Graduate Student Member, IEEE) received the B.S. degree in mechanical design, manufacturing, and automation from the Wuhan University of Technology, Wuhan, China, in 2018. He is currently pursuing the Ph.D. degree with the State Key Laboratory of Digital Manufacturing Equipment and Technology, Huazhong University of Science and Technology, Wuhan.

His current research interests include intelligent manufacturing, defects detection, image processing, and deep learning.

Haiming Yao (Student Member, IEEE) received the B.S. degree from the School of Mechanical Science and Engineering, Huazhong University of Science and Technology, Wuhan, China, in 2022. He is pursuing the Ph.D. degree with the Department of Precision Instrument, Tsinghua University, Beijing, China.

His research interests include deep learning, edge intelligence, and machine vision.

Lixin Tang received the B.S. and M.S. degrees from the Huazhong University of Science and Technology, Wuhan, China, in 1989 and 1992, respectively, and the Ph.D. degree from the University of Tsukuba, Tsukuba, Japan, in 2002.

He is currently an Associate Professor with the School of Mechanical Science and Engineering, Huazhong University of Science and Technology. His current research interests include image process, computer vision, and pattern recognition.

Wenyong Yu (Member, IEEE) received the M.S. and Ph.D. degrees from the Huazhong University of Science and Technology, Wuhan, China, in 1999 and 2004, respectively.

He is currently an Associate Professor with the School of Mechanical Science and Engineering, Huazhong University of Science and Technology. His research interests include machine vision, intelligent control, and image processing.

Bin Li (Member, IEEE) received the B.S., M.S., and Ph.D. degrees in mechanical engineering from the Huazhong University of Science and Technology, Wuhan, China, in 1982, 1989, and 2006, respectively.

He is currently a Professor with the School of Mechanical Science and Engineering, Huazhong University of Science and Technology. His current research interests include intelligent manufacturing and computer numerical control (CNC) machine tools.