AFSC: Adaptive Fourier Space Compression for Anomaly Detection

Haote Xu, Yunlong Zhang, Xiaolu Chen, Student Member, IEEE, Changxing Jing, Liyan Sun, Yue Huang, Member, IEEE, and Xinghao Ding

Abstract—The primary challenge faced by reconstruction-based anomaly detection (AD) methods is that neural networks exhibit strong generalization, resulting in a high probability and accuracy of anomaly reconstruction. Several existing methods attempt to alleviate this problem by randomly masking partial image regions and reconstructing the image from partial inpaintings. However, local masking in spatial space is not guaranteed to remove anomalous regions during the testing phase and poses the risk of normal regions being inaccurately reconstructed. Hence, we explore an approach to compress the global information of the image while ensuring the loss of partial anomaly information renders it difficult to reconstruct. Inspired by the fact that each Fourier coefficient contains global information of the image, we propose an adaptive Fourier space compression (AFSC) method. Specifically, the Fourier coefficients of the input image are sparsely sampled by binary masks obtained from the AFSC module (AFSCm). In AFSCm, the masks are jointly optimized with the reconstruction network subject to sparsity constraint. The learned masks are forced to selectively retain part of the global information that is favourable to recovering normal images. In addition, we introduce an efficient Fourier convolution module that enables the network to accurately reconstruct normal regions under conditions of losing partial information. Experimental results on three benchmarks of industrial scenarios demonstrate our method (without external prior) achieves competitive results compared with recent methods.

Index Terms—Anomaly detection (AD), Fourier space, global context, reconstruction-based anomaly detection.

I. INTRODUCTION

Anomaly detection (AD) is an essential task with diverse applications across various industrial inspections, including the detection of surface defects and structural defects, etc [1]. The majority of current approaches facilitate community development by leverage techniques, such as self-supervised learning [2], [3], feature distance metric [4], [5], [6], and image augmentation [7], [8]. The reconstruction-based methods, as one of the mainstream approaches, enable AD by calculating the differences between the detected sample and its reconstruction. It is predicated on the assumption that the reconstruction network cannot reconstruct anomalous regions not encountered during the training phase.

However, the excellent generalization capability of neural networks leads to accurate reconstruction of certain anomalies when directly fed with original images [9]. Such successful reconstruction of these anomalous region results in false negatives for partial anomalous regions, as illustrated in the first row of Fig. 1. This phenomenon contradicts the initial fundamental assumption, leading to the failure of the reconstruction-based AD methods. Zavrtanik et al. [9] mitigated the aforementioned issue by increasing the difficulty of successfully reconstructing anomalous regions, which is achieved by introducing random masks in the input images. Nevertheless, the implementation of random masking in the spatial space cannot ensure the effective masking of anomalous regions. On the contrary, normal regions will be difficult to reconstruct with high fidelity and be erroneously detected as anomalous due to the substantial losing information, as shown in the second row of Fig. 1. According to these findings, we explore how to accurately reconstruct normal regions while impeding the reconstruction of anomalous areas with losing information.

It is a well-known property of Fourier space that each coefficient contains global signals about the image [10], [11]. We, therefore, utilize binary masks to compress global information in Fourier space ensuring the removal of partial anomalous region signals. With the suppression of the anomalous region signals, the difficulty of reconstructing these regions can be increased. Ideally, for the normal regions that also experienced partial signal loss, the network with the objective function of reconstructing the normal image can efficiently reconstruct them during the inference phase by utilizing the remaining Fourier
coefficients. However, the random masks in the Fourier space introduce the potential for less-than-desirable reconstruction results in normal regions, as illustrated in the third row of Fig. 1.

In this work, we propose an adaptive Fourier space compression (AFSC) for AD, which consists of two modules, i.e., an adaptive fourier space compression module (AFSCm) and an efficient Fourier convolution module (EFCm). Specifically, in the AFSCm, we design a strategy that jointly optimizes the Fourier space masks and the reconstruction network under the sparsity constraint, which is inspired by Monte Carlo strategies [12]. In this way, the network is endowed with the ability to obtain masks as sparse as possible and achieve reconstruction of normal regions. The EFCm is introduced to further avoid the reconstruction of normal regions with low fidelity. The existing literature suggests that combining local context in shallow layers and global context is beneficial for most visual tasks [11], [13], [14]. However, the traditional convolutional neural networks adopts the architecture of stacking multiple layers with small kernels (e.g., 3 × 3 in residual network (ResNet) [15]) to increase the receptive field, resulting in the global context being visible only at deep layers. This causes low efficacy in the information combination of two distant contexts in the network. Therefore, we design the EFCm by considering again the property of the Fourier space masks and the reconstruction network under the sparsity constraint, which is inspired by Monte Carlo strategies [12].

In this work, we propose an adaptive Fourier space compression (AFSC) for AD, which consists of two modules, i.e., an adaptive fourier space compression module (AFSCm) and an efficient Fourier convolution module (EFCm). Specifically, in the AFSCm, we design a strategy that jointly optimizes the Fourier space masks and the reconstruction network under the sparsity constraint, which is inspired by Monte Carlo strategies [12]. In this way, the network is endowed with the ability to obtain masks as sparse as possible and achieve reconstruction of normal regions. The EFCm is introduced to further avoid the reconstruction of normal regions with low fidelity. The existing literature suggests that combining local context in shallow layers and global context is beneficial for most visual tasks [11], [13], [14]. However, the traditional convolutional neural networks adopts the architecture of stacking multiple layers with small kernels (e.g., 3 × 3 in residual network (ResNet) [15]) to increase the receptive field, resulting in the global context being visible only at deep layers. This causes low efficacy in the information combination of two distant contexts in the network. Therefore, we design the EFCm by considering again the property of the Fourier space. EFCm learns features containing global context by transforming shallow features into Fourier space. Simultaneously, it utilizes an attention mechanism to facilitate combining these transformed features with shallow features. Integrating EFCm into the shallow layers of the reconstruction network enables the network to efficiently combine local context and global context, thereby enhancing reconstruction capability.

The main contributions are the following.

1) We alleviate the problems of some anomalous regions being successfully reconstructed for reconstruction-based AD methods by compressing information in Fourier space.

2) We propose AFSC, which obtains masks jointly optimized with the network for Fourier space compression, enabling high-fidelity reconstruction of normal regions while allowing the masks as sparse as possible.

3) As a reconstruction-based AD method, the AFSC achieves comparable results to existing methods on three datasets, employing only U-Net as the backbone network without external prior (external training data or pretraining models).

II. RELATED WORK

Existing AD methods can be broadly divided into feature-based approaches and reconstruction-based approaches, according to the performing AD space.

A. Feature-Based Approaches

The feature space-based approaches perform AD in the feature space by comparing the difference between the feature distribution of anomaly samples and normal samples. Classical methods treat support vector machine [16] as one-class classification for AD, e.g., support vector data description (SVDD) [17]. Ruff et al. [18] proposed a deep SVDD method that combines deep learning with SVDD. However, deep SVDD can only determine whether an image is anomalous or not without the ability to localize anomalous regions. To this end, Yi et al. [19] introduced patch SVDD, which generates multiple clustering centers in the training phase by clustering the features of normal sample patches. Recently, Bergmann et al. [4] proposed an AD method based on knowledge distillation. Specifically, AD is achieved by assuming a large regression error in the representation of the anomaly sample between the student networks and the teacher network. Similarly, Salehi et al. [5], in order to improve the network’s ability to extract discriminative features, try to distinguish outlier features based on multiple intermediate layer features of the student and teacher networks. Besides, PatchCore [20] extracts patch-level features and uses nearest-neighbor methods to classify patches and images as anomalies. Similarly, Shape-guided [21] utilizes normal training samples to build dual memory banks for the 3-D structure and 2-D RGB features, which are used for AD in 3-D scenarios.

B. Reconstruction-Based Approaches

The reconstruction-based approaches for AD perform AD by comparing the differences between the reconstructed image and the input image in spatial space. A typical approach is training an autoencoder based on L2-distance or structural similarity metric (SSIM) [22] as the loss function [23]. In the test phase, the pixel-level difference between the image and its reconstruction is calculated by the loss to detect anomalies. It is based on the assumption that the unavailability of anomalies in the training process will cause anomaly reconstruction to fail. However, neural networks have strong generalization capability, leading to frequent success in anomalous reconstructions. Models based on generative adversarial networks (GAN) [24] can avoid the issue. Such as the anomaly detection with generative adversarial networks (AnoGAN) [25] by searching for the nearest representation of the input image in the latent space constructed from the normal image and decoding it. However, the inference
process of AnoGAN is time-consuming as it needs to iteratively search for the nearest representation in latent space for the input. To mitigate this issue, fast-anoGAN (f-AnoGAN) [26] effectively improves inference speed by introducing a discriminator. In general, the normal region reconstruction accuracy of the GAN-based methods is not as good as the regular reconstruction methods. Zavrtanik et al. [7] proposed discriminatively trained reconstruction anomaly embedding model (DRAEM), which constructs anomaly samples via auxiliary datasets to obtain a reconstruction network with recovery capability for out-of-distribution anomalies. Furthermore, DRAEM utilizes a discriminator to learn the distance function between the original image and its reconstruction for AD. Considering that the methods of image-level anomaly synthesis are prone to failure when facing near-in-distribution anomalies, Zavrtanik et al. [8] proposed to generate feature-level anomalies, allowing the controlled generation of near-in-distribution anomalies to improve the generalization capability of the model. Reconstruction by inpainting for visual anomaly detection (RIAD) [9] attempts to randomly mask the image to lose anomalous regions, then repair the masked regions to pseudonormal regions by inpainting. Unlike existing approaches, we design a simple method, AFSC, which is mainly tailored to the problem of certain anomalous regions being successfully reconstructed in reconstruction-based AD methods.

III. METHOD

Given a training set of normal samples $\mathcal{X} = \{x_i\}_{i=1}^n$, where $x_i$ denotes a normal sample in $\mathcal{X}$. The goal of reconstruction-based AD is to learn a model $f(\cdot; \theta)$ that effectively detects anomalies by accurately reconstructing normal samples and failing to reconstruct anomaly samples during the inference phase. Nevertheless, the neural networks generalize well even to certain anomalous regions and reconstruct them sufficiently well.

In this article, we suggest compressing the Fourier space signal of the input image in order to destroy its global information, thereby preventing the reconstruction model from reconstructing the anomalous regions of the input image. Specifically, we present an AFSCm to generate masks that can be jointly optimized with the network to reach a tradeoff between sparse regularization of Fourier space and reconstruction of normal images. Furthermore, to tackle the challenge of accurately reconstructing normal regions affected by global information destroyed, the EFCm is proposed to unlock the potential of the network to combine local context and global context at shallower layers. The overall architecture of AFSC will be introduced in the following sections, along with its main components, namely the AFSCm and the EFCm. We show the architecture of the AFSC in Fig. 2.

A. Overview of the AFSC

For a single channel image $x \in \mathbb{R}^{1 \times H \times W}$ that is used as input to the network, its Fourier transformation $\mathcal{F}(x)$ is expressed as

$$\mathcal{F}(x)(u, v) = \sum_{h=0}^{H-1} \sum_{w=0}^{W-1} x(h, w)e^{-j2\pi \frac{h u}{H} + \frac{v w}{W}}.$$  

As for an RGB image $x \in \mathbb{R}^{3 \times H \times W}$, the Fourier transformation can be calculated independently for each channel. For the sake of description, we take the example of a single-channel image below. Since the values in $\mathcal{F}(x)(u, v)$ are complex numbers, we divide them into the real part $R(x)(u, v)$ and the imaginary part $I(x)(u, v)$.
where the binary mask \( M \) for compressing partial Fourier space signals is obtained by AFSCm. The real and imaginary parts of the compressed Fourier space are formulated as follows:

\[
\hat{R}(x) = R(x) \odot M \tag{2}
\]

\[
\hat{I}(x) = I(x) \odot M \tag{3}
\]

where the binary mask \( M \) is obtained by joint optimization with the network. The details of \( M \) are described in the Section III-B. The masked real and imaginary parts are combined to form a new Fourier representation

\[
\hat{F}(\hat{x}) = \left[ \hat{R}^2(x) + \hat{I}^2(x) \right]^{1/2} \tag{4}
\]

which applies with the inverse Fourier transformation to obtain a Fourier space-compressed image

\[
\hat{x} = \mathcal{F}^{-1} [\hat{F}(\hat{x})(u,v)] \tag{5}
\]

where \( \mathcal{F}^{-1} \) defines the inverse Fourier transformation. Finally, \( \hat{x} \) is used as the input to the reconstruction network. The total loss function of the network as follows:

\[
\mathcal{L} = \mathcal{L}_2 + \mathcal{L}_{SSIM} + \mathcal{L}_{MSGMS} + \alpha \mathcal{L}_{color} + \beta \mathcal{L}_M \tag{6}
\]

where the first three terms refer to the reconstruction losses utilized in [9]. These losses consist of pixel-wise \( L_2 \) loss, SSIM loss \( \mathcal{L}_{SSIM} \) [22], and multiscale gradient magnitude similarity (MSGMS) loss \( \mathcal{L}_{MSGMS} \). Furthermore, we incorporate the color difference loss, which corresponds to the fourth term in (6) in order to enhance the accuracy of the reconstruction. Specifically, we employ the function proposed in [27] to calculate color difference between the reconstructed image and the original image. This is achieved by transforming both into the commission on illumination lab (CIELAB) color space and calculating the differences between their respective components as the loss. The last term \( \beta \mathcal{L}_M = \beta \| M \|_1 \) allows \( M \) to satisfy the sparsity property. The hyperparameters \( \beta \) of larger values tend to obtain more sparse masks compared to \( \beta \) of smaller values.

### B. Adaptive Fourier Space Compression Module

We assume that each element \( m_i \) of \( M \) is obtained by sampling in the Bernoulli distribution with probability \( p_i \). Sampling of the Bernoulli distribution is achieved by reparameterising \( m_i = \Delta \{ u_i \leq p_i \} \) in AFSCm, where \( \Delta \{ \cdot \} \) is a function that maps a Boolean-true input to 1 or 0, and \( u_i \) is a random variable sampled from the [0,1] uniform distribution. However, the optimization process is hindered due to the nondifferentiability of \( \Delta \{ \cdot \} \). Hence, we relax the function \( \Delta \{ \cdot \} \) by approximating it as the differentiable sigmoid function \( \sigma(\cdot) \).

First, we sample a tensor \( \omega \in \mathbb{R}^{H \times W} \) from the uniform [0,1] distribution and then map it to an optimizable variable \( \hat{\omega} \) via the inverse function of the sigmoid function \( \sigma(\cdot) \). Thus, the probability map \( p = \sigma(\hat{\omega}) \) can be built by \( \hat{\omega} \). And the \( m_i = \Delta \{ u_i \leq p_i \} \) can be approximated in the following optimizable form:

\[
M = \sigma \left( \sigma(\hat{\omega}) - u \right), \quad u \sim U(0,1). \tag{7}
\]

The abovementioned steps are depicted in left part of Fig. 2. Ultimately, the \( M \) is obtained by joint optimization with the network, enabling the network to reconstruct the normal images while compressing the Fourier space.

### C. Efficient Fourier Convolution Module

The architecture of EFCm is shown in right part of Fig. 2. EFCm consists of two pathways: a local pathway that performs regular convolution on input features, and a global pathway that operates to transform features into the Fourier space. Each pathway can capture information from different receptive fields. Meanwhile, the internal exchange of information between the two pathways will occur to achieve information complementary. Then, the outputs of the two pathways are concatenated and sent into the attention mechanism module to obtain the final output. Formally, let \( \phi_{\text{EFCm}}^\text{in} \) indicates the input feature map of EFCm. First, \( \phi_{\text{EFCm}}^\text{in} \) is split into two parts of the same dimension: the input \( \phi_{\text{local}}^\text{in} \) of the local pathway and the input \( \phi_{\text{global}}^\text{in} \) of the global pathway. Subsequently, \( \phi_{\text{local}}^\text{in} \) and \( \phi_{\text{global}}^\text{in} \) are separately

---

**Algorithm 1: Pseudocode of Fourier Transformer**

**Input:** feature \( \phi_{\text{in}} \in \mathbb{R}^{B \times C \times W \times H} \# B \): batchsize, \( C \): dimension of channel, \( H, W \): spatial size of feature

**Output:** feature \( \phi_{\text{out}} \in \mathbb{R}^{B \times C \times W \times H} \)

# Fourier Transformer

**def FT(\phi_{\text{in}}):**

\( x_o = \text{ReLU}(\text{BN}(\text{Conv}(\phi_{\text{in}}))) \# [B, C, H, W] \)

\( z_o = \text{FTu}(x_o) \# [B, \frac{C}{2}, H, W] \)

\( x_s = \text{Concatenate}(\text{Split}(x_o[, :, C/8, 2], H/2, \text{dim} = -2), \text{dim} = 1) \# [B, \frac{C}{2}, H, W] \)

\( x_s = \text{Concatenate}(\text{Split}(x_s, W/2, \text{dim} = -1), \text{dim} = 1) \# [B, C, H, W] \)

\( z_s = \text{FTu}(x_s) \# [B, \frac{C}{2}, H, W] \)

# Spatially shift and replicate 4 copies of \( z_s \)

\( z_s = \text{shift}(z_s, \text{repeat}(1, 1, 2, 2)) \# [B, C, H, W] \)

\( \phi_{\text{out}} = \text{Conv}(x_o + z_o + z_s) \# [B, C, H, W] \)

**return** \( \phi_{\text{out}} \)

# FT unit

**def FTu(x):**

\( y_r, y_i = \text{FFT}(x) \# y_r, y_i \) is the real/imaginary part of the results of Fast Fourier Transform, respectively. \( \text{FTF}() \) means Computing the 2-dimensional discrete Fourier transform of input.

\( y = \text{Concatenate}([y_r, y_i], \text{dim} = 1) \)

\( y = \text{ReLU}(\text{BN}(\text{Conv}(y))) \)

\( y_r, y_i = \text{Split}(y, \text{dim} = 1) \)

\( z = \text{iFFT}(y_r, y_i) \# \text{iFFT}() \) means Computing the 2-dimensional discrete inverse Fourier transform of input.

**return** \( z \)
introduced into the four branches of two pathways, which can be described by the following formulas:

\[ \phi^{g \rightarrow g} = f_{EFCm}^{\text{local}}(\phi_{\text{in}}^{g \rightarrow g}) \]
\[ \phi^{g \rightarrow l} = f_{EFCm}^{\text{global}}(\phi_{\text{in}}^{g \rightarrow l}) \]
\[ \phi^{l \rightarrow g} = f_{EFCm}^{\text{local}}(\phi_{\text{in}}^{l \rightarrow g}) \]
\[ \phi^{l \rightarrow l} = f_{EFCm}^{\text{global}}(\phi_{\text{in}}^{l \rightarrow l}) \]  \hspace{1cm} (8)

where \( f_{EFCm}^{\text{local}}(\cdot) \) denotes local-to-local branches, \( f_{EFCm}^{g \rightarrow g}(\cdot) \) and \( f_{EFCm}^{g \rightarrow l}(\cdot) \) refer to local-to-global and global-to-local branches utilized for intrainformation exchange. All of these branches employ standard 3 × 3 convolution modules. The \( f_{EFCm}^{l \rightarrow l}(\cdot) \) signifies the use of a Fourier transformer (FT) in the global-to-global branch, as detailed in Algorithm 1. The FT realizes that transforming original spatial features into Fourier space for global contextual feature learning. Following the practice in [11], we split and concatenate the input features, yielding smaller features \( x_s \) for capturing and propagating semilocal information in FT.

The outputs of the local and global pathway are formulated as \( \phi^{\text{out}}_{\text{local}} \) and \( \phi^{\text{out}}_{\text{global}} \), respectively. The \( \phi^{\text{out}}_{\text{global}} \) is designed to capture global context, while \( \phi^{\text{out}}_{\text{local}} \) is expected to learn information of the local context. These outputs are ultimately concatenated to obtain \( \phi^{\text{out}}_{\text{EFCm}} \). However, the importance of different channel features in \( \phi^{\text{out}}_{\text{EFCm}} \) for the reconstruction task is distinct. Therefore, the channel attention mechanism can be employed to explicitly model the interdependencies between feature channels, enhancing important features while suppressing task-irrelevant ones. We introduce an approach that efficiently facilitates local interchannel interactions through one-dimensional convolution. Specifically, we employ one-dimensional convolution to each channel with its neighbouring channels to efficiently calculate the channel weights \( s_c \), after applying the global average pooling as used in SENet [28]. The final output of EFCm is obtained through the following:

\[ \phi^{\text{out}}_{\text{EFCm}} = f_{\text{scale}}(\phi^{\text{out}}_{\text{EFCm}}, s_c) \]  \hspace{1cm} (9)

where \( f_{\text{scale}}(\cdot, \cdot) \) refers to channel-wise multiplication between the scalar \( s_c \) and the feature map \( \phi^{\text{out}}_{\text{EFCm}} \).

D. Theoretical Analysis

Both modules are proposed based on the property that each coefficient in Fourier space contains partial global information of the spatial space. In AFSCm, the effect of the Fourier space mask \( M \) on the original spatial signal will be analyzed by comparing the signal difference between \( \hat{x} \) and \( x \) in Section III-A. The inverse Fourier transform \( \mathcal{F}^{-1} \) of the masked Fourier space signal \( \mathcal{F}(\hat{x})(u, v) \) in (5) yields spatial space the signal that can be expressed in the following form:

\[ \hat{x}(h, w) = \frac{1}{HW} \sum_{u=0}^{H-1} \sum_{v=0}^{W-1} [\mathcal{F}(x)(u, v) \circ M(u, v)] \cdot e^{j2\pi\left(\frac{h}{H} + \frac{w}{W}\right)} \]  \hspace{1cm} (10)

where \( \circ \) represents the composition of functions in (2), (3), and (4). For comparison, the inverse Fourier transform of the original signal \( x \) is

\[ x(h, w) = \frac{1}{HW} \sum_{u=0}^{H-1} \sum_{v=0}^{W-1} \mathcal{F}(x)(u, v) \cdot e^{j2\pi\left(\frac{h}{H} + \frac{w}{W}\right)}. \]  \hspace{1cm} (11)

Clearly, the mask \( M(u, v) \) modifies the Fourier space coefficients \( \mathcal{F}(x)(u, v) \), which in turn affects every point \( (h, w) \) in the spatial space. Since the Fourier transform is a global transformation, each Fourier coefficient \( \mathcal{F}(x)(u, v) \) is derived from the entire spatial signal \( x(h, w) \). Therefore, any change in the Fourier coefficients affects the entire spatial space signal.

Besides, according to Parseval’s theorem [29], the energy in the spatial space is equivalent to the energy in the Fourier space

\[ \sum_{h=0}^{H-1} \sum_{w=0}^{W-1} |x(h, w)|^2 = \frac{1}{HW} \sum_{u=0}^{H-1} \sum_{v=0}^{W-1} |\mathcal{F}(x)(u, v)|^2. \]  \hspace{1cm} (12)

After applying the mask, the energy relationship becomes

\[ \sum_{h=0}^{H-1} \sum_{w=0}^{W-1} |\hat{x}(m, n)|^2 = \frac{1}{HW} \sum_{u=0}^{H-1} \sum_{v=0}^{W-1} |\mathcal{F}(x)(u, v) \circ M(u, v)|^2. \]  \hspace{1cm} (13)

Since the mask \( M \) reduces or eliminates certain frequency components

\[ \sum_{u=0}^{H-1} \sum_{v=0}^{W-1} |\mathcal{F}(x)(u, v) \circ M(u, v)|^2 \leq \sum_{u=0}^{H-1} \sum_{v=0}^{W-1} |\mathcal{F}(x)(u, v)|^2. \]  \hspace{1cm} (14)

This indicates that the masking operation reduces the energy in the Fourier space, which must also manifest as a change in the overall energy distribution and information in the spatial space.

The abovementioned analysis proves that applying the masking operation in the Fourier space affects the entire spatial space signal. As the Fourier transform is a global transformation, any change in the Fourier space coefficients impacts all points in the spatial space. Moreover, Parseval’s theorem shows that energy loss in the frequency domain is reflected in the spatial domain. Thus, AFSCm’s operations in the Fourier space of the input image affect the global information in the spatial space, hence ensuring that the loss of anomaly information prevents anomalies from being reconstructed with high fidelity.

Similarly, the Fourier transform of a feature followed by a convolution operation in the global path of the EFCm can be seen as combining the contextual information of the feature globally. Contrary to convolution in the original feature space, the convolution operation in the Fourier domain improves the ability to capture long-range dependencies, although it weakens the ability to extract local information. Finally, EFCm performs convolution operations in both the features’ original space and the Fourier space to extract local and global contextual information, respectively. By fusing these features, EFCm captures hierarchical features, thereby overcoming the limitation of shallow layers that can only extract local information.
IV. EXPERIMENTS

To demonstrate the effectiveness of our method, extensive experiments on industrial inspection AD were performed. We compare AFSC with ten recent AD techniques, including feature-based methods, such as multiresolution knowledge distillation (MKD) [5], PatchCore [20], PyramidFlow [30], and self-supervised normalizing flows (SSNF) [31], reconstruction-based methods, such as DRAEM [7], dual subspace re-projection network (DSR) [8], EdgRec [27], and RIAD [9], and 3-D AD methods, such as Shape-guided [21] and back to the feature (BTF) [32]. The methods MKD, PatchCore, PyramidFlow, SSNF, DRAEM, DSR, and EdgRec incorporate external priors, specifically Visual Geometry Group 16 (VGG-16), Wide Residual Networks 50 (WRes50), Residual Networks 18 (Res18), Class-Attention in Image Transformers (CaiT), Describable Textures Dataset (DTD), ImageNet, and DTD, respectively. In the following, we present the relevant datasets, metrics, implementation details, and experimental results in industrial inspection scenarios.

A. Datasets and Metrics

We evaluate our method on visual anomaly (VisA) [2], MVTec AD [1], and MVTec-3D AD [33] datasets. The VisA dataset contains 12 industrial objects and consists of 10 821 images in total with 9621 normal and 1200 anomalous samples. The MVTec AD dataset is also a benchmark for AD, which collected 3535 colour images from 10 object categories and 5 texture categories. The MVTec-3D AD [3] dataset consists of 10 classes totaling 2656 training samples and 1137 test samples. Each sample stores RGB information and position information in two 3-channel tensors, respectively. Note that there are only normal samples in the training set, while the testing set contains both normal and anomaly samples. In the experiment, the image sizes for VisA, MVTec AD, and MVTec-3D AD are resized to $256 \times 256$. The pixel values are normalized to the range $[0, 1]$.

In this article, we evaluate AFSC in terms of image-level AD and pixel-level anomaly localization. Following [20], the area under the receiver operator curve (AUROC) is used to evaluate the performance of image-level AD. The localization performance is evaluated by employing the pixel-level AUROC and per-region-overlap (PRO).

B. Implementation Details

The network architecture of AFSC is based on the U-Net framework. EFCm is introduced only before the second and third standard convolutional blocks of the network to combining local context and global context in the shallow layers with a small additional computational burden. To mitigate overfitting, augmented normal images are used to generate artificial anomaly samples instead of the external dataset in [7] during the training phase.

The network is trained on two datasets for both 800 epochs, with the batchsize of 2. We use the Adam optimizer with an initial learning rate of $1e^{-4}$ and reduce the learning rate by a factor of 0.2 at the 640 and 720 epoch. For the hyperparameter $\alpha$, the setting from [27] is adopted, which is consistently set to $1e^{-2}$ in all experiments. For the hyperparameter $\beta$, we consistently set its value to $1e^{-6}$ among all experimental scenarios. Notably, the variable $M$, derived through the AFSCm, takes on the form of a soft mask, wherein each element $M_{ij} \in (0, 1)$. Consequently, upon reaching the 400th epoch, we binarize it through a voting mechanism. Specifically, we conduct a tally if the element in $M_{ij}$ exceeds the threshold $u_{ij}$ during each iteration spanning epochs 396 to 400, wherein $u_{ij}$ represents an element of $U$ in (7). The values of $u_{ij}$ are randomly sampled from a uniform distribution and resampled in each iteration. After the completion of the 400th epoch, we perform a tally of the vote counts. If the tally exceeds half of $5(epochs) \times k$, where $k$ represents the number of iterations per epoch, we assign 1 to the corresponding element; otherwise, it is set to 0.

C. Comparison With Existing Methods

1) Results on VisA: We report the results of image-level AD and pix-level anomaly localization for VisA in Table I. The average Img-AUROC results across all categories indicate that despite the absence of external priors, our method still achieves the best detection performance. AFSC outperforms the best-performing feature-based methods (i.e., PatchCore [20]) by 2.8%. And in the reconstruction-based methods, compared to the second place (i.e., EdgRec [27]) our method improves by 1.8%. In terms of anomaly localization, AFSC maintains strong anomaly localization performance compared to current advanced methods. It achieves the Pix-AUROC and PRO average scores over all categories with 98.3% and 90.9%. Although our method is lower than the previous top performance in terms of average PRO scores, it gets the best performance in average pixel-level AUROC. Some examples of qualitative results on the VisA are shown in the upper part of Fig. 3.

2) Results on MVTec AD: Table II reports the detection and localization performance on MVTec. The average score over all classes shows that our method achieves the second-best performance among all compared methods after PatchCore. Specifically, the proposed AFSC reach the average AUROC (98.4%) in all categories. In terms of the localization task, our method achieves the highest average AUROC score. In addition, among all the reconstruction-based approaches, AFSC exceeds the second place (EdgRec [40]) by 1.0% and 1.5% at the pixel level AUROC and PRO metrics, respectively. Since each category of the MVTec AD contains different types of anomalies, this excellent performance in AUROC metrics significantly demonstrates that the AFSC is effective in detecting and localizing different types of anomalous regions. Some examples of qualitative results on the MVTec AD are shown in the lower part of Fig. 3.

3) Results on MVTec-3D AD: In Table III, both 3-D methods demonstrated performance superior to all 2-D methods. The fact that the 2-D methods, which do not utilize the 3-D modality of the samples and can only perform AD based on the RGB modality, yield such results is not surprising. The 2-D methods perform poorly in image-level AD tasks. Our method achieved the best performance among the 2-D methods with an AUROC...
### TABLE I
RESULTS OF THE IMG-AUROC/PIX-AUROC/PRO (IN %) METRICS FOR AD AND LOCALIZATION PERFORMANCE ON VISA

| Category | Feature-based methods | Reconstruction-based methods | VGG-16 | WR-255 | ResNet | CAT | DRAEM | DSR | EdgRec | RIAD | Ours |
|----------|-----------------------|-----------------------------|--------|--------|--------|-----|-------|-----|-------|------|------|
| Pcl1     | MKD [5]               | PatchCore [20]             | 91.199/4.272 | 98.248/3.256 | 96.285/6.262 | 93.898/3.878 | 83.487/2.126 | 97.614/35.875 | 49.002/5.176 | 32.004/4.375 |
| Pcl2     | MKD [5]               | PatchCore [20]             | 91.199/4.272 | 98.248/3.256 | 96.285/6.262 | 93.898/3.878 | 83.487/2.126 | 97.614/35.875 | 49.002/5.176 | 32.004/4.375 |
| Pcl3     | MKD [5]               | PatchCore [20]             | 91.199/4.272 | 98.248/3.256 | 96.285/6.262 | 93.898/3.878 | 83.487/2.126 | 97.614/35.875 | 49.002/5.176 | 32.004/4.375 |
| Pcl4     | MKD [5]               | PatchCore [20]             | 91.199/4.272 | 98.248/3.256 | 96.285/6.262 | 93.898/3.878 | 83.487/2.126 | 97.614/35.875 | 49.002/5.176 | 32.004/4.375 |

The best and second-best performance is highlighted in bold and underlined, respectively.

### TABLE II
RESULTS OF THE IMG-AUROC/PIX-AUROC/PRO (IN %) METRICS FOR AD AND LOCALIZATION PERFORMANCE ON MVTEC AD

| Category | Feature-based methods | Reconstruction-based methods | VGG-16 | WR-255 | ResNet | CAT | DRAEM | DSR | EdgRec | RIAD | Ours |
|----------|-----------------------|-----------------------------|--------|--------|--------|-----|-------|-----|-------|------|------|
| Carpet   | MKD [5]               | PatchCore [20]             | 91.199/4.272 | 98.248/3.256 | 96.285/6.262 | 93.898/3.878 | 83.487/2.126 | 97.614/35.875 | 49.002/5.176 | 32.004/4.375 |
| Grid     | MKD [5]               | PatchCore [20]             | 91.199/4.272 | 98.248/3.256 | 96.285/6.262 | 93.898/3.878 | 83.487/2.126 | 97.614/35.875 | 49.002/5.176 | 32.004/4.375 |
| Leather  | MKD [5]               | PatchCore [20]             | 91.199/4.272 | 98.248/3.256 | 96.285/6.262 | 93.898/3.878 | 83.487/2.126 | 97.614/35.875 | 49.002/5.176 | 32.004/4.375 |
| Tile     | MKD [5]               | PatchCore [20]             | 91.199/4.272 | 98.248/3.256 | 96.285/6.262 | 93.898/3.878 | 83.487/2.126 | 97.614/35.875 | 49.002/5.176 | 32.004/4.375 |
| Wood     | MKD [5]               | PatchCore [20]             | 91.199/4.272 | 98.248/3.256 | 96.285/6.262 | 93.898/3.878 | 83.487/2.126 | 97.614/35.875 | 49.002/5.176 | 32.004/4.375 |
| Pill     | MKD [5]               | PatchCore [20]             | 91.199/4.272 | 98.248/3.256 | 96.285/6.262 | 93.898/3.878 | 83.487/2.126 | 97.614/35.875 | 49.002/5.176 | 32.004/4.375 |
| Transistor| MKD [5]               | PatchCore [20]             | 91.199/4.272 | 98.248/3.256 | 96.285/6.262 | 93.898/3.878 | 83.487/2.126 | 97.614/35.875 | 49.002/5.176 | 32.004/4.375 |
| Cable    | MKD [5]               | PatchCore [20]             | 91.199/4.272 | 98.248/3.256 | 96.285/6.262 | 93.898/3.878 | 83.487/2.126 | 97.614/35.875 | 49.002/5.176 | 32.004/4.375 |
| Zipper   | MKD [5]               | PatchCore [20]             | 91.199/4.272 | 98.248/3.256 | 96.285/6.262 | 93.898/3.878 | 83.487/2.126 | 97.614/35.875 | 49.002/5.176 | 32.004/4.375 |
| Toothbrush| MKD [5]               | PatchCore [20]             | 91.199/4.272 | 98.248/3.256 | 96.285/6.262 | 93.898/3.878 | 83.487/2.126 | 97.614/35.875 | 49.002/5.176 | 32.004/4.375 |
| Metal nut| MKD [5]               | PatchCore [20]             | 91.199/4.272 | 98.248/3.256 | 96.285/6.262 | 93.898/3.878 | 83.487/2.126 | 97.614/35.875 | 49.002/5.176 | 32.004/4.375 |
| Harzmat  | MKD [5]               | PatchCore [20]             | 91.199/4.272 | 98.248/3.256 | 96.285/6.262 | 93.898/3.878 | 83.487/2.126 | 97.614/35.875 | 49.002/5.176 | 32.004/4.375 |
| Screw    | MKD [5]               | PatchCore [20]             | 91.199/4.272 | 98.248/3.256 | 96.285/6.262 | 93.898/3.878 | 83.487/2.126 | 97.614/35.875 | 49.002/5.176 | 32.004/4.375 |
| Capsule  | MKD [5]               | PatchCore [20]             | 91.199/4.272 | 98.248/3.256 | 96.285/6.262 | 93.898/3.878 | 83.487/2.126 | 97.614/35.875 | 49.002/5.176 | 32.004/4.375 |
| Bottle   | MKD [5]               | PatchCore [20]             | 91.199/4.272 | 98.248/3.256 | 96.285/6.262 | 93.898/3.878 | 83.487/2.126 | 97.614/35.875 | 49.002/5.176 | 32.004/4.375 |
| Avg.     | MKD [5]               | PatchCore [20]             | 91.199/4.272 | 98.248/3.256 | 96.285/6.262 | 93.898/3.878 | 83.487/2.126 | 97.614/35.875 | 49.002/5.176 | 32.004/4.375 |

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
score of 84.6%. Besides, our method, even when using only a single modality, achieves results that are not far behind the second place, which is remarkable. Unlike image-level detection, in pixel-level anomaly localization tasks, most 2-D methods have shown ideal results.

4) Computational Efficiency: In this section, we provide a detailed analysis of method complexity in terms of the number of parameters (Params) and inference time. As shown in Table IV, the frames per second (FPS), floating-point operations (FLOPs), and the number of Params for all methods are reported. Notably, all methods are tested on devices equipped with a 3080ti GPU. Based on the results reported in Table IV and the overall performance across the three datasets, it is evident that our method achieves high computational efficiency while maintaining detection accuracy. In addition, it can be observed that there is no apparent positive correlation between the FPS of the aforementioned methods and their FLOPs and Params. This phenomenon occurs because hyperparameters only indicate the size of the network. In practice, some methods may involve multiple computations due to the use of multiple input images. For instance, Riad [9] performs multiple masking and reconstruction operations on input images, which hinders efficient inference. Furthermore, certain FLOPs are not captured by the current FLOPs measurement tools. For example, Shape-Guided [21] extracts features from all training images using a pretrained network and then performs a patching operation, storing 10% of the core patch features as a coreset. During testing, the test images undergo a similar feature extraction and patching process. However, the subsequent KNN comparison of all patch features with the coreset is not accounted for by FLOPs measurement tools. Shape-Guided [21] also employs the similar operation. The +86.1 (+16.9) reported in the parameters (Params) in Table IV represents the parameter count of its core subset of shape-guided (PatchCore).

Overall, reconstruction-based methods, except for AFSC, show less stability across different datasets compared to feature-based methods. The feature-based methods typically leverage pretrained networks for fine-tuning or do not require the training process (e.g., PatchCore), resulting in relatively stable performance differences across datasets. In contrast, reconstruction-based methods often need to introduce external datasets for augmentation during the training process. The introduced images lack generalization, resulting in significant performance differences in the network across various datasets. However, this issue can be mitigated. Our method does not rely on external data and compresses global image information, making anomalies difficult to restore. This contributes to the method conduct stable performance across different datasets. Nevertheless, AFSC and comparison methods perform inferior to MVTec AD on the VisA dataset. This is attributed to the fact that VisA incorporates multiobject classes, thereby increasing the detection difficulty. It is reasonable that the detection performance on 3-D dataset declines when only single-modality information is utilized. However, our method consistently achieves promising results in image-level AD in both datasets relative to the comparison methods. The phenomenon is attributed to the ability of AFSCm, as a module of reconstructing images with losing global information, to make anomalous regions destroyed and hardly reconstructed, differently from [7], [8], [9], [27]. In addition, EFCm enables the network to efficiently combine local context and global context at the shallow layers, thus reducing the probability of false positives by achieving accurate reconstruction of normal regions.

D. Ablation Studies

1) Validity of Modules: Tables V–VIII demonstrate the effectiveness of the proposed AFSCm and EFCm, where STD

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
TABLE VI
ABLA TION RESULTS OF THE IMG-AUROC/PIX-AUROC/PRO (IN %) METRICS FOR AD AND LOCALIZA TION PERFORMANCE ON VisA

| Category       | U-Net | Random-FC | ASFCm | ASFC-DCT | ASFC | DWT |
|----------------|-------|------------|-------|----------|------|-----|
| Tns            | 75.59 | 75.59      | 75.59 | 75.59    | 75.59| 75.59|
| Resi           | 68.68 | 68.68      | 68.68 | 68.68    | 68.68| 68.68|
| Polar          | 57.59 | 57.59      | 57.59 | 57.59    | 57.59| 57.59|
| Prin           | 57.59 | 57.59      | 57.59 | 57.59    | 57.59| 57.59|
| Eufi           | 57.59 | 57.59      | 57.59 | 57.59    | 57.59| 57.59|
| Cocc           | 57.59 | 57.59      | 57.59 | 57.59    | 57.59| 57.59|
| Corr           | 57.59 | 57.59      | 57.59 | 57.59    | 57.59| 57.59|
| Addi           | 57.59 | 57.59      | 57.59 | 57.59    | 57.59| 57.59|
| Npli           | 57.59 | 57.59      | 57.59 | 57.59    | 57.59| 57.59|
| Bold           | 57.59 | 57.59      | 57.59 | 57.59    | 57.59| 57.59|

In this section, we conduct experimental analysis by replacing the discrete Fourier transform (DFT) in the proposed method with the discrete wavelet transform (DWT) and discrete cosine transform (DCT). The experimental results show that DFT achieves the best performance in all datasets. The reason for the abovementioned phenomenon is that DFT can perform a global frequency analysis of the entire image signal, accurately representing all frequency components across different frequency bands. In contrast, DCT mainly focuses on the low-frequency part and has weaker representation capabilities for high-frequency details and global frequency components. DWT decomposes the signal into multiple sub-bands containing different frequency and directional information, resulting in significant differences between sub-bands of different images. Since ASFC learns the unified masks for all training images, using DFT provides the ASFC architecture with stronger generalization capabilities compared to the others.

2) Impact of Hyperparameter \( \beta \): In order to investigate the effect of mask sparsity on the performance in ASFCm, we seek to change the scale of the hyperparameter \( \beta \) and observe the variation tendency in performance. As shown in Table VI, the poor performance is obtained when \( \beta \) is set with a large value. Since the masks for the Fourier space are too sparse to reconstruct the image, the model's performance culminated in its optimum at \( \beta = 1 \).

denotes Standard Deviation of the methods. We build the baseline through vanilla U-Net, which has a poor performance in both datasets. The overall performance is improved by incorporating ASFCm to obtain the adaptive Fourier space masks instead of the random masks. The difficulty of reconstructing anomalous regions is increased by the ASFCm, which compresses the global information of the image. Simultaneously, ASFCm preserves the capacity for the reconstruction of normal regions through joint optimization with the reconstruction network. Finally, the ASFC model is obtained by incorporating EFCm, which achieves (3.6%, 1.7%, 11.3%) image-level AUROC, (1.9%, 0.6%, 2.6%) pixel-level AUROC, and (9.4%, 5%, 8.3%) PRO performance improvement at MVTec AD/VisA/ MVTec 3-D AD compared to the baseline. Fig. 4 illustrates the qualitative reconstruction results of the methods in ablation Studies. The reconstruction results of ASFC (last column) are more similar to normal samples (first column) compared to other methods, indicating the robust image normality recovery ability of our method.

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
V. CONCLUSION

In this article, we argue that the problem of certain anomalous regions being reconstructed in reconstruction-based AD methods can be alleviated by compressing the Fourier space information. We propose an AFSCm to obtain masks that are jointly optimized with the network, enabling the network to adaptively select the Fourier spatial coefficients for accurate reconstruction of normal images. Furthermore, the EFSCm is introduced to facilitate the combination of global and local contexts at shallow layers within the network, thereby enhancing the accuracy of normal region reconstruction. Overall, we propose to compress information in Fourier space based on the physical properties of the Fourier coefficient for reconstruction-based AD methods, which presents an alternative viewpoint for AD. Nevertheless, this work still presents certain limitations. First, AFSC is specifically designed for AD tasks in traditional 2-D scenarios. While it demonstrates some advantages over existing 2-D methods in 3-D scenarios, it still falls short when compared to dedicated 3-D methods. Second, although AFSC surpasses existing methods in terms of inference time, it is at a disadvantage regarding the number of Params. Therefore, extending AFSC more effectively to 3-D scenarios, by utilizing the position information rather than relying solely on RGB information from images, is a direction that requires further exploration. In addition, designing more lightweight networks while maintaining performance is another consideration for our future research. More importantly, exploring how to better leverage information in the Fourier space is a key direction for our future work.

REFERENCES

[1] 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., 2019, pp. 9892–9899.

[2] Y. Zou, J. Jeong, L. Pemula, D. Zhang, and O. Dabeer, “Spot-the-difference self-supervised pre-training for anomaly detection and segmentation,” in Proc. Eur. Conf. Comput. Vis., 2022, pp. 392–408.

[3] J. Jiang et al., “Masked Swin transformer U-Net for industrial anomaly detection,” IEEE Trans. Ind. Informat., vol. 19, no. 2, pp. 2200–2209, Feb. 2023.

[4] P. Bergmann, M. Fauser, D. Sattlegger, and C. Steger, “Uninformed students: Student-teacher anomaly detection with discriminative latent embeddings,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2020, pp. 4183–4192.

[5] M. Salehi, N. Sadjadi, S. Baselizadeh, M. H. Rohban, and H. R. Rabiee, “Multiresolution knowledge distillation for anomaly detection,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2021, pp. 14902–14912.

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

[7] V. Zavratnik, M. Kristan, and D. Škocj, “DRAEM—the discriminatively trained reconstruction embeddings for surface anomaly detection,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., 2021, pp. 8330–8339.

[8] V. Zavratnik, M. Kristan, and D. Škocj, “DSR—a dual subspace reconstruction network for surface anomaly detection,” in Proc. Eur. Conf. Comput. Vis., 2022, pp. 539–554.

[9] V. Zavratnik, M. Kristan, and D. Škocj, “Reconstruction by inpainting for visual anomaly detection,” Pattern Recognit., vol. 112, 2021, Art. no. 107706.

[10] Y. Katznelson, An Introduction to Harmonic Analysis. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[11] J. Lin, C. Gan, and S. Han, “TSM: Temporal shift module for efficient video understanding,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., 2019, pp. 7083–7093.

[12] D. P. Kingma and M. Welling, “Auto-encoding variational Bayes,” 2013, arXiv:1312.6114.

[13] L. Zhu et al., “Bidirectional feature pyramid network with recurrent attention residual modules for shadow detection,” in Proc. Eur. Conf. Comput. Vis., 2018, pp. 121–136.

[14] 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.

[15] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 770–778.

[16] C. Cortes and V. Vapnik, “Support-vector networks,” Mach. Learn., vol. 20, no. 3, pp. 273–297, 1995.

[17] D. M. Tax and R. P. Duin, “Support vector data description,” Mach. Learn., vol. 54, no. 1, pp. 45–66, 2004.

[18] L. Ruff et al., “Deep one-class classification,” in Proc. Int. Conf. Mach. Learn., 2019, pp. 4393–4402.

[19] J. Yi and S. Yoon, “Patch SVDD: Patch-level SVDD for anomaly detection and segmentation,” in Proc. Asian Conf. Comput. Vis., 2020, pp. 375–390.

[20] K. Roth, L. Pemula, J. Zepeda, B. Schölkopf, T. Brox, and P. Gehler, “Towards total recall in industrial anomaly detection,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2022, pp. 14318–14328.

[21] Y.-M. Chu, C. Liu, T.-I. Hsieh, H.-T. Chen, and T.-L. Liu, “Shape-guided dual-memory learning for 3D anomaly detection,” in Proc. 40th Int. Conf. Mach. Learn., 2020, pp. 6185–6194.

[22] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, “Image quality assessment: From error visibility to structural similarity,” IEEE Trans. Image Process., vol. 13, no. 4, pp. 600–612, Apr. 2004.

[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] I. Goodfellow et al., “Generative adversarial networks,” Commun. ACM, vol. 63, no. 11, pp. 139–144, 2020.

[25] T. Schlegl, P. Seeböck, S. M. Waldstein, U. Schmidt-Erfurth, and G. Langs, “Unsupervised anomaly detection with generative adversarial networks to guide marker discovery,” in Proc. Int. Conf. Inf. Process. Med. Imag., 2017, pp. 146–157.

[26] 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, 2019.

[27] T. Liu, B. Li, Z. Zhao, X. Du, B. Jiang, and L. Geng, “Reconstruction from edge image combined with color and gradient difference for industrial surface anomaly detection,” 2022, arXiv:2210.14485.

[28] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2018, pp. 7132–7141.

[29] B. Jähne, Digital Image Processing, Berlin, Germany: Springer, 2005.

[30] J. Lei, X. Hu, Y. Wang, and D. Liu, “PyramidFlow: High-resolution defect contrastive localization using pyramid normalizing flow,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2023, pp. 14143–14152.

[31] L.-L. Chiu and S.-H. Lai, “Self-supervised normalizing flows for image anomaly detection and localization,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops, 2023, pp. 2926–2935.

[32] E. Horwitz and Y. Hoshen, “Back to the feature: Classical 3D features are (almost) all you need for 3D anomaly detection,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2023, pp. 2967–2976.

[33] P. Bergmann, X. Jin, D. Sattlegger, and C. Steger, “The MVTEC 3D-AD dataset for unsupervised 3D anomaly detection and localization,” 2021, arXiv:2112.09045.

Haote Xu was born in Wenzhou, China, in 1995. He is currently working toward the Ph.D. degree in information and signal processing with the School of Informatics, Xiamen University, Xiamen, China.

His research interests include machine learning, anomaly detection, and medical image analysis.

Dr. Xu is also a Reviewer for some journals such as IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, etc.
Yunlong Zhang received the B.E. degree in applied physics from Southwestern Jiaotong University, Chengdu, China, in 2018 and the M.E. degree in information and communication engineering from Xiamen University, Xiamen, China, in 2021. He is currently working toward the Ph.D. degree in computer science and technology with the School of Engineering, Westlake University, Hangzhou, China. His research interests include computer vision and medical image analysis.

Xiaolu Chen (Student Member, IEEE) received the B.E. degree in electronic information engineering from Shanghai Maritime University, Shanghai, China, in 2022. She is currently working toward the M.E. degree in communication engineering with the School of Informatics, Xiamen University, Xiamen, China. Her research interests include anomaly detection and image processing.

Changxing Jing received the Ph.D. degree in signal and information processing from the School of Informatics, Xiamen University, Xiamen, China, in 2023. His research interests include federated learning, representation learning, computer vision, and machine learning.

Dr. Jing is currently a Reviewer for journals and conferences such as IEEE CYBER SCIENCE AND TECHNOLOGY CONGRESS, and WIRELESS COMMUNICATIONS AND MOBILE COMPUTING, etc.

Liyan Sun received the B.E. degree in communication engineering from Zhengzhou University, Zhengzhou, China, in 2014 and the Ph.D. degree in signal and information processing from the School of Informatics, Xiamen University, Xiamen, China, in 2021. He is currently a Postdoctoral Researcher with the Department of Radiation Oncology and Stanford University School of Medicine, Stanford University, Stanford, CA, USA. His research interests mainly focus on machine learning and medical image analysis.

Yue Huang (Member, IEEE) received the B.S. degree in communication engineering from the Department of Electrical Engineering, Xiamen University, Xiamen, China, in 2005 and the Ph.D. degree in bioengineering from the Department of Biomedical Engineering, Tsinghua University, Beijing, China, in 2010. She is currently an Professor with the School of Informatics, Xiamen University. Her main research interests include sparse signal representation and machine learning. Her main research interests include image processing, sparse signal representation, and machine learning.

Xinghao Ding was born in Hefei, China, in 1977. He received the B.S. degree in precision instrument and Ph.D. degree in precision instrument and machinery from the Department of Precision Instruments, Hefei University of Technology, Hefei, China, in 1998 and 2003, respectively. He was a Postdoctoral Researcher with the Department of Electrical and Computer Engineering, Duke University, Durham, NC, USA, from 2009 to 2011. Since 2011, he has been a Professor with the School of Informatics, Xiamen University, Xiamen, China. His main research interests include machine learning, deep learning, computer vision, and signal processing.