Transformer-based Multi-scale Reconstruction Network for Defect Detection of Infrared Images

Changyun Wei, Hui Han, Zhichao Wu, Yu Xia, and Ze Ji*

***Abstract*—Bottle packaging is extensively used in manufacturing, and inspecting aluminum foil sealing during filling is crucial for ensuring product quality. Traditional Machine vision methods based on supervised learning require extensive annotated data, but the scarcity of defective samples hampers the effectiveness of these methods. To address this challenge, unsupervised learning methods have emerged. Despite their potential, these methods often struggle to accurately learn the distribution of normal samples, resulting in higher rates of false positives and negatives. This paper proposes an unsupervised learning-based approach for anomaly detection in infrared images. Specifically, we construct a Transformer-based multi-scale image reconstruction network (TMIRN) that includes a feature extraction module, a feature fusion module, a reconstruction module, a discriminator network, and an anomaly scoring module. By effectively combining Transformer and CNN techniques, the proposed network excels at capturing both global and local semantic information. Its multi-scale structure accurately localizes defects of varying sizes and combines image-level and feature-level anomaly scores to mitigate the impact of non-uniform distribution and noise. Experimental results on the infrared image dataset for aluminum foil sealing demonstrate high accuracy in anomaly detection and localization. Furthermore, on the industrial MVTec AD dataset, our TMIRN exhibits superior generalization and detection compared to state-of-the-art reconstruction networks.**

***Index Terms*—Industrial defect detection, infrared images, unsupervised learning, vision transformer.**

I. Introduction

In the field of manufacturing, bottle packaging has widespread adoption across various industries including food, pharmaceuticals, cosmetics, and others. A critical aspect of bottle packaging involves the utilization of aluminum foil sealing technology, which holds significant importance in upholding product quality [1]. Factors such as temperature variations and the quality of the aluminum foil sheets directly influence the efficacy of the sealing process, thereby impacting the overall quality of the final product. However, with the rapid advancements in industrial automation, traditional inspection methods reliant on sampling, such as the water pressure and air pressure methods, fall short in meeting the requisites of fully automated mass production lines used in filling operations. Consequently, there exists an urgent necessity for a swift and efficient method capable of identifying defects in aluminum foil sealing.

This work was supported in part by the National Natural Science Foundation of China under Grant 52371275. (Corresponding authors: Ze Ji (jiz1@cardiff.ac.uk).

Changyun Wei, Hui Han, Zhichao Wu and Yu Xia are with the College of Mechanical and Electrical Engineering, Hohai University, China.

Ze Ji is with the School of Engineering, Cardiff University, Cardiff CF24 3AA, United Kingdom.

With the rapid development of computer and imaging technology, machine vision inspection methods have become widely utilized in industrial defect detection. These methods offer non-invasive, efficient, safe, and reliable means of assessment [2]. Infrared imaging technology, known for its affordability, simplicity, and broad scanning range, is particularly suitable for detecting internal defects in industrial products like materials and machinery [3], [4]. Fig. 1 showcases the infrared images of aluminum foil sealing in various cases. From a machine vision perspective, the infrared images can be used to identify sealing defects after the aluminum foil sealing process.

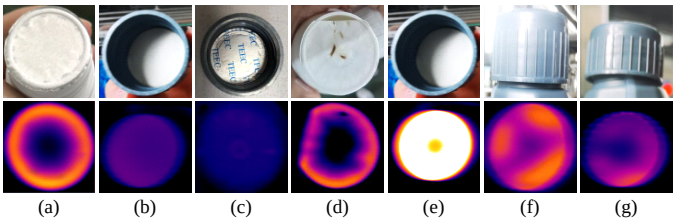


Fig. 1. Infrared images in aluminum foil sealing. (a) normal, (b) no aluminum foil, (c) unsealed, (d) nicked, (e) reversed, (f) loose cap, (g) crooked cap.

The majority of conventional machine vision techniques rely on extracting predefined artificial features from infrared images or implementing threshold-based measures. These methods include classical Principal Component Thermography (PCT) and its extensions [5], [6], template matching, as well as techniques such as wavelet integral alternating sparse dictionary matrix decomposition [7], among others. While these methods can efficiently detect defects under specific conditions, they impose stringent demands on the operational environment. Their effectiveness often hinges on the necessity for targeted parameter adjustments, making them less adaptable to diverse scenarios. Moreover, their generalizability across varied conditions is limited. Additionally, employing traditional methods necessitates a substantial reservoir of expert experience to yield reliable outcomes.

In recent years, the advancements in deep learning have notably revolutionized image processing, allowing for the acquisition of intricate semantic features from images with enhanced robustness and generalizability [8]. This progress has facilitated the integration of intelligent techniques into industrial nondestructive testing based on infrared imaging. Presently, deep learning methods utilized in industrial defect detection are principally categorized into supervised and unsupervised learning approaches. Within the domain of industrial defect detection, supervised learning predominantly leverages

convolutional neural networks such as YOLO, R-CNN, and others to extract meaningful features from images [9], [10], [11]. However, in authentic industrial settings, supervised learning encounters the following inevitable limitations.

1) The continuous upgrading of industrial production lines leads to the difficulty of obtaining defect samples.
2) The manual labeling of datasets incurs substantial labor costs.
3) The collected defect samples may not comprehensively cover all potential defect cases.

Hence, the adoption of unsupervised learning for anomaly detection has garnered attention in defect detection processes. Models trained solely on normal samples exhibit promise in identifying and pinpointing defective instances within datasets. To the best of our knowledge, anomaly detection models are rarely developed to infrared image-based defect detection.

Presently, two primary approaches dominate anomaly detection methodologies: image reconstruction-based and feature embedding-based methods. Image reconstruction-based models, such as autoencoders and generative adversarial networks (GANs), are trained on defect-free samples to hinder the accurate reconstruction of defective samples [12], [13], [14], [15], [16], [17], [18], [19], [20]. When an image with defects is input, the trained model reconstructs the normal portions accurately but often fails to do so for the anomalous portions, using pixel-by-pixel error as the anomaly score. To enhance reconstruction quality, the work [12] employs the structural similarity index (SSIM) [21] as a loss function, and Samet et al. [15] present the Skip-GANomaly network associated with skip connections inspired by U-Net [22]. However, over-generalization can lead to a decrease in detection accuracy due to the false reconstruction of the anomalous features. Methods like memory storage [13] and feature clustering [14] have been employed to address this issue. Yang et al. [14] adopt feature clustering in the MS-FCAE model to restrict spatial distribution, and Zavrtanik et al. [23] employ inpainting techniques to improve local information reconstruction. Despite these advancements, unstable local feature reconstruction and the neglect of deep semantic features at the feature level still affect defect detection accuracy. Additionally, these models are sensitive to inhomogeneous backgrounds and noise in infrared images.

Feature embedding-based methods primarily use pre-trained networks like ResNet18 and ResNet50 to extract generalized features from defect-free images. Techniques such as normalized streaming, feature memory banks, and K-means clustering are then used to model the normal distribution of these features [24], [25], [26], [27], [28]. Anomalies are identified by comparing extracted features with the normal feature distribution. For instance, PaDiM [24] uses a pre-trained ResNet to extract features and models their distribution with a Gaussian distribution, calculating anomaly scores using the Mahalanobis distance. PatchCore [26] uses features from a pre-trained ResNet and builds a memory bank to calculate anomaly scores through nearest neighbor search. However, significant challenges arise in accurately extracting relevant features using pre-trained models due to the category differences between industrial infrared images and natural images. These disparities can lead to feature mismatches. To achieve this, SimpleNet [28] combines synthetic and embedding-based methods by employing a feature adapter to reduce domain bias and introducing noise into the feature space to generate anomalies, thereby producing in a more tightly bounded normal feature space. Moreover, the variable nature of contours in infrared images complicates object localization, and CNN-based networks struggle to capture global semantic features, hindering precise segmentation for large or distant defects.

As an attention-based feature extraction network, the Transformer architecture, initially used in natural language processing, has extended into computer vision. Dosovitskiy et al. [29] introduce the Vision Transformer (ViT) for image classification, showing robust learning across diverse datasets. Compared to conventional CNNs, ViT excels at capturing extensive global dependencies within images. However, Transformer-based networks have limited applications in anomaly detection. Pirnay et al. [30] employ the Transformer module for restoring masked image patches, focusing on normal image reconstruction. Chen et al. [25] utilize a pre-trained ResNet18 for multi-scale fusion feature extraction, feeding these features into the U-Transformer network for feature reconstruction and defect detection based on reconstruction error. Similarly, Tao et al. [31] use ViT for feature extraction, integrating a hybrid structure of Transformer and pyramid architecture with an anomaly estimation module for fine-grained defect localization. While the Vision Transformer is popular in image processing, it tends to prioritize global semantic information, potentially overlooking local details during reconstruction. Additionally, its high memory consumption and computational cost pose practical challenges.

To address the aforementioned challenges, the paper proposes a novel Transformer-based Multiscale Image Reconstruction Network (TMIRN). The TMIRN consists of several key modules: a multiscale Transformer feature extraction module, a multiscale attention feature fusion module, a multiscale Transformer reconstruction module, a discriminator network and an anomaly score module. By adopting a GAN structure, the model combines CNN and Transformer architectures for encoding and decoding processes, incorporating an inductive bias into the Transformer to effectively exploit both global and local semantic information. Moreover, the model demonstrates robust performance even with limited datasets. The approach employs upsampling and convolutional structures to replace traditional deconvolution layers, reducing reconstruction artifacts. To enable accurate defect localization at varying scales, the model embraces a multi-scale design, integrating image-level multi-scale anomaly scores with feature-level anomaly scores. This strategy mitigates the impact of inhomogeneous backgrounds and noise prevalent in infrared images. To further enhance the network's reconstruction capabilities, the paper introduces a blend of skip connections and bottleneck structures to effectively suppress anomalous features while ensuring accurate reconstruction of normal images. The main contributions of this paper are summarized as follows:

1) Introducing TMIRN with a GAN structure for unsupervised learning, enhancing infrared image reconstruction

and addressing sample scarcity and annotation costs in aluminum foil sealing.
2) Developing multi-scale Transformer feature extraction and reconstruction modules that leverage global and local semantic information, enhancing model reconstruction capabilities.
3) Integrating image-level and feature-level multi-scale anomaly scores for precise defect detection and localization, achieving high precision on a self-collected infrared image dataset and robust performance on the MVTec AD dataset.

II. Background

Aluminum foil sealing technology is a highly efficient sealing method that operates on the principle of electromagnetic induction heating [32]. This technique involves the sophisticated design of layered structures and is predominantly utilized in the food, pharmaceutical, and cosmetic industries to enhance the safety and shelf life of products. The core component of this sealing process is the aluminum foil liner, whose quality is crucial to the sealing performance. The aluminum foil liner consists of several layers: a polymer layer, an aluminum foil layer, a wax bond layer, and a pulp board, as illustrated in Fig. 2. During the sealing process, the aluminum

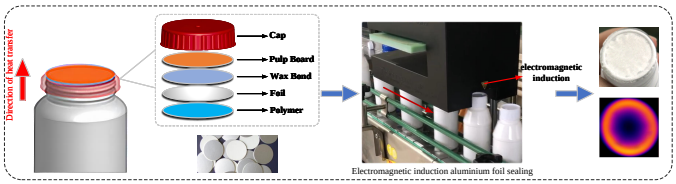


Fig. 2. Schematic diagram of the aluminum foil sealing technology.

foil liner is first embedded into the bottle cap, exposing the aluminum foil layer. The cap is then screwed onto the bottle. At this stage, the electromagnetic induction equipment generates instant high heat on the aluminum foil layer, causing the polymer and wax bond layers to melt. The polymer layer bonds firmly with the container's opening, while the wax bond layer ensures that the aluminum foil layer separates from the pulp board. Following the electromagnetic induction sealing, the aluminum foil layer retains a high temperature, conducting heat upwards. Due to the adhesive properties of the polymer and the increased compressive force applied when the cap is screwed on, the temperature at the bottle's edge is rapidly transmitted to the surface of the cap. As a result, thermal imaging captured by an infrared camera shows a continuous ring-shaped pattern. During the cooling process, the polymer layer solidifies, forming a secure seal. This sealing method effectively prevents the contents from leaking and being exposed to moisture.

III. Methods

The proposed network in this paper adopts the unsupervised learning methodology, using only normal samples for training, and it differs from the basic autoencoder architectures such as AE-SSIM [12] and DRAEM [33]. To be specific, we enhance the detail perception by combining transformer modules and employing multiscale feature fusion. To address the common issue in reconstruction networks, the proposed Efficient Channel Attention (ECA) module can suppress defect features during information propagation. Additionally, the combination of pixel-level and feature-level anomaly scores can effectively mitigate noise impact during reconstruction, thereby improving the accuracy of anomaly detection and localization.

A. Main Structure of TMIRN

The core components of TMIRN include the multiscale transformer feature extraction (MTFE) module, the multiscale attention feature fusion (MAFF) module, the multiscale transformer decoder (MTD), a discriminator network, and an anomaly score module. Together, the MTFE and MAFF modules form the multiscale encoder within this framework. The overall structure is shown in Fig. 3.

During the training phase, the enhanced multiscale images $\boldsymbol{I}_{in,x}$ are fed into the MTFE module to extract local and global semantic features. These features are then fused by the MAFF module into a single representation $\boldsymbol{Z}$. To enhance reconstruction details and improve information flow, $\boldsymbol{Z}$ is passed through a bottleneck structure with skip connections before being input into the MTD to reconstruct the corresponding multiscale image $\boldsymbol{I}_{out,x}$. To improve the quality and realism of the reconstructed infrared images, a discriminator network, sharing the same structure as the multiscale encoder, is incorporated to work with the generator in adversarial training. This network captures both global structures and fine details, enabling the model to assess image granularity at multiple levels, thereby improving the realism of the reconstructed images.

During the inference phase, residuals between the multiscale inputs $\boldsymbol{I}_{in,x}$ and outputs $\boldsymbol{I}_{out,x}$ are computed to generate residual maps. These maps are then fused and denoised to create the image-level anomaly map $\boldsymbol{S}_i$. Meanwhile, features from the intermediate layers of the encoder and decoder are extracted, resized, and used to compute additional residuals. These residuals are smoothed by the anomaly score smoothing module, producing the feature-level anomaly map $\boldsymbol{S}_f$. Finally, the image-level and feature-level maps are combined to produce the overall anomaly score map $\boldsymbol{S}$.

B. Multiscale Transformer Feature Extraction (MTFE)

To proficiently identify sealing defects across multiple scales, the MTFE incorporates a diverse architecture, as depicted in Fig. 4. This structure is designed to address the complexities of sealing defects, which can vary in size and appearance across different scales. In order to extract varying levels of feature information from input images at large, medium, and small scales, the MTFE captures both global contextual information and fine-grained details that are essential for accurate defect identification.

Specifically, the module integrates the Vision Transformer to enhance the model's emphasis on global contextual information. This enables the network to understand the broader context of the image, aiding in the detection of anomalies spanning multiple scales. To preserve spatial details, edges,

and fine-grained features during the feature extraction stage, a residual convolution module is introduced before the Vision Transformer module. This supplementary module facilitates downsampling and local feature extraction, thereby improving

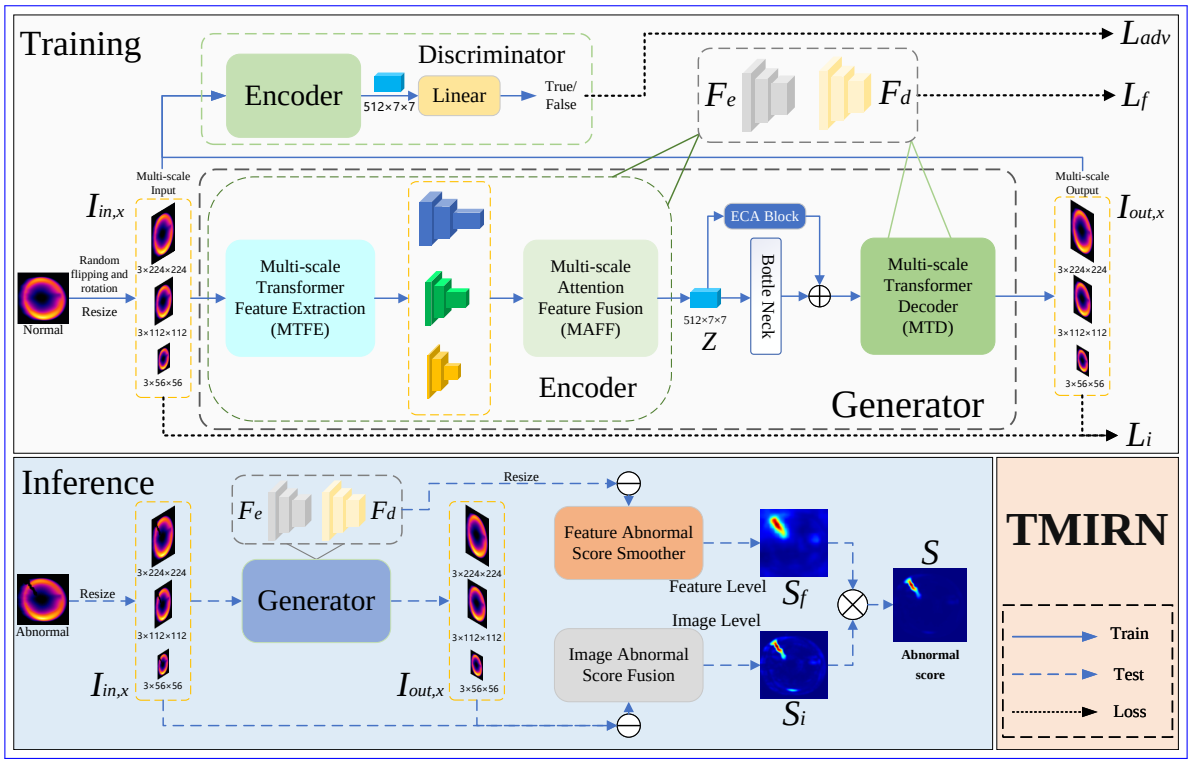


Fig. 3. The proposed TMIRN framework for defect detection and localization of infrared images. The generator includes an encoder, composed of the Multiscale Transformer Feature Extraction (MTFE) module and the Multiscale Attention Feature Fusion (MAFF) module, and the Multiscale Transformer Decoder (MTD). The discriminator shares the same structure as the encoder. The anomaly scoring module consists of both image-level and feature-level components.

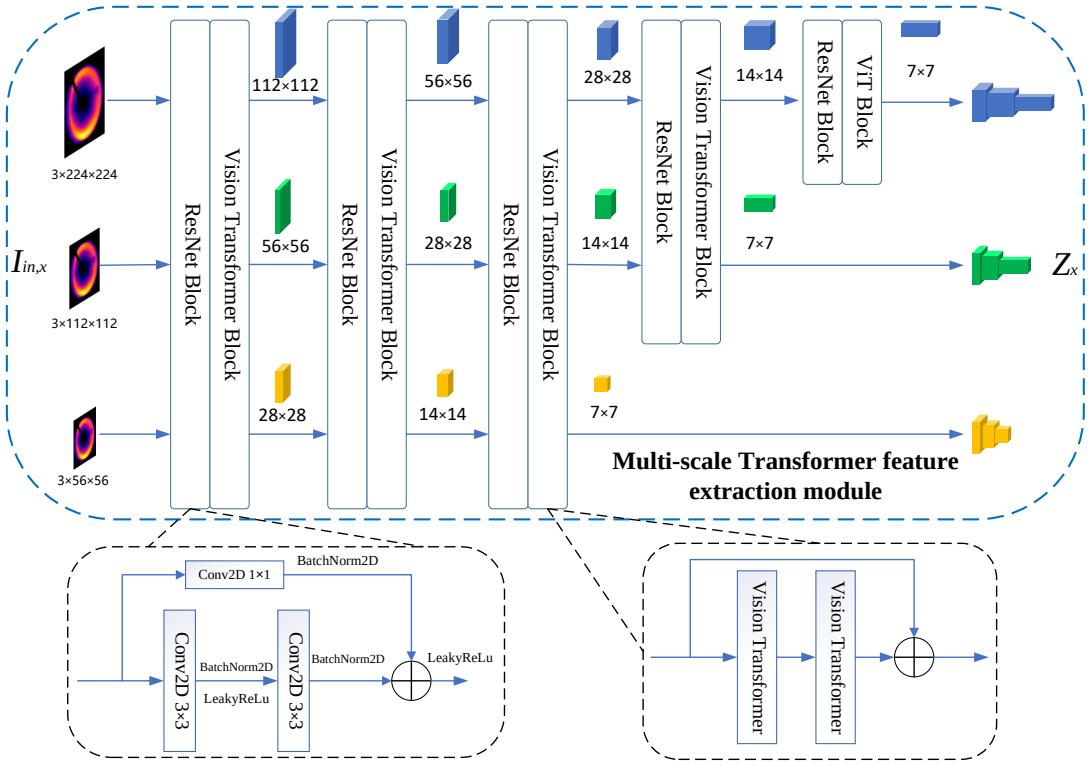


Fig. 4. The structure of the Multiscale Transformer Feature Extraction (MTFE) module.

the model's ability to discern subtle defects. Each layer of the MTFE comprises a residual convolution module and a Vision Transformer module. This structure, illustrated in Fig. 4, leverages the strengths of both convolutional and transformer-based architectures. The residual convolution module, consisting of two 3×3 convolutions and a skip connection with a 1×1 convolution, helps in capturing detailed local features, while the BatchNorm and LeakyReLU activations expedite model convergence and enhance training stability.

Assume that we have the input feature map $\boldsymbol{T} \in \boldsymbol{R}^{H \times W \times C}$,

$$\widehat{\boldsymbol{T}} = \mathrm{LR}(\mathrm{Conv}_{3\times3}(\mathrm{LR}(\mathrm{Conv}_{3\times3}(\boldsymbol{T}))) + \mathrm{Conv}_{1\times1}(\boldsymbol{T})),$$
$$\widehat{\boldsymbol{T}} \in \boldsymbol{R}^{\frac{H}{2} \times \frac{W}{2} \times 2C}, \tag{1}$$

where LR(.) represents the LeakyRelu activation function, and H, W, C denote the height, width and dimension of the feature map, respectively. The number of output channels in the convolutional layer is set to $2C$.

The Vision Transformer module comprises two lightweight transformers and a skip connection. The improved Vision Transformer structure, inspired by PVTv2 [34], is depicted in Fig. 5. This paper adopts a lightweight multi-head self-

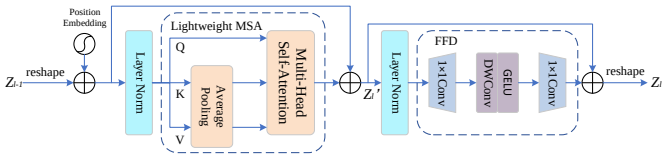


Fig. 5. The improved structure of the Vision Transformer.

attention mechanism and an enhanced feed-forward structure to effectively reduce the computational cost of the transformer. The lightweight multi-head attention mechanism serves to reduce computational overhead by employing downsampling techniques on K and V using a $k \times k$ mean pooling operation to obtain smaller features K' and V'. This mechanism is formulated as follows:

$$K' = \mathrm{AvgPool}(K) \in \boldsymbol{R}^{\frac{n}{k^2} \times d_k} \tag{2}$$

$$V' = \mathrm{AvgPool}(V) \in \boldsymbol{R}^{\frac{n}{k^2} \times d_v} \tag{3}$$

$$\mathrm{LMSA} = \mathrm{Softmax}(\frac{QK'^T}{\sqrt{d_k}})V', \tag{4}$$

where $Q, K \in \boldsymbol{R}^{n \times d_k}$, $V \in \boldsymbol{R}^{n \times d_v}$, k denotes the kernel size for mean pooling and $n = H \times W$ means the number of patches.

The output of the residual convolution module is initially mapped to sequence space before being fed into the lightweight multi-head self-attention (LMSA) and feed-forward network (FFD) for subsequent processing. This approach improves computational efficiency by using a 1×1 convolution instead of full connectivity, thereby reducing complexity while maintaining effectiveness. Layer Norm and skip connections are integrated to stabilize training and enhance gradient flow, contributing to the overall robustness and performance of the model in anomaly detection tasks.

$$Z_l{}' = \mathrm{LMSA}(\mathrm{LN}(Z_{l-1} + E_{pos})) + (Z_{l-1} + E_{pos}) \tag{5}$$

$$\mathrm{FFD} = \mathrm{Conv}(\mathrm{DWConv}(\mathrm{Conv}(X))) \tag{6}$$

$$Z_l = \mathrm{FFD}(\mathrm{LN}(Z_l{}')) + Z_l{}', \tag{7}$$

where LN(.) indicates the layer normalization, Z_l means the embedding features of the l-th layer, and E_{pos} as the positional embedding is equal in dimension size to Z_{l-1}.

The MTFE conducts feature extraction on input images of various scales to obtain multiscale feature mappings of corresponding sizes. Specifically, if the input multiscale images are represented as $\boldsymbol{I}_{in,x}$ (x=l, m, s), the resulting multiscale features are denoted as $Z_x = \{z_{x,low}, z_{x,mid}, z_{x,high}\}$.

$$\{z_{x,low}, z_{x,mid}, z_{x,high}\} = \{f_{n,x}(\boldsymbol{I}_{in,x}), f_{n+1,x}(\boldsymbol{I}_{in,x}),$$
$$f_{n+2,x}(\boldsymbol{I}_{in,x})\}, x = l, m, s \tag{8}$$

where $f_{n,x}(.)$ indicates the output of the n-th feature extraction module.

C. Multiscale Attention Feature Fusion (MAFF)

As the multiscale feature mappings derived from MTFE showcase notable disparities in spatial characteristics, semantic information, and size, the MAFF module is utilized to execute cross-scale fusion of this intricate information.

In order to enhance the feature representation capacity and empower the model to concentrate on more pertinent channel information during the fusion process while suppressing irrelevant feature channels, the ECA module is integrated within the feature fusion module. A visual representation of this structure is depicted in Fig. 6.

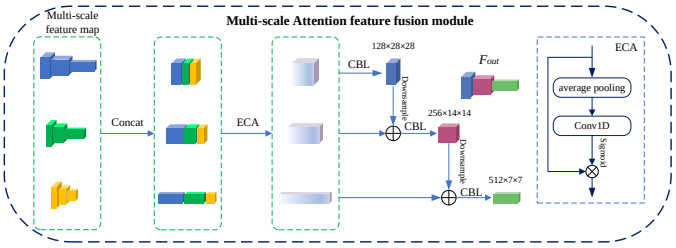


Fig. 6. The structure of the Multiscale Attention Feature Fusion (MAFF) module.

Initially, the MTFE module generates multiscale features, which are rearranged and concatenated to capture diverse contextual information. This design enhances the model's ability to detect anomalies of varying sizes and complexities by fusing local and global features effectively. The ECA module assigns varying weights to channels based on their importance in capturing relevant features. This adaptive channel weighting mechanism improves anomaly detection performance by focusing on discriminative features and reducing noise. The structure of the ECA module is depicted in Fig. 6, where a sequence of weights mirroring the dimensions of the features is obtained through mean-pooling and one-dimensional convolutional operations. This sequence of weights is multiplied with the original features to obtain the channel-weighted feature blocks. Subsequently, the feature blocks assigned with channel weights are input into the feature fusion pyramid, and the fusion of different semantic features is performed through

downsampling and concatenation convolution operations to obtain three different scales of fused features. The detailed dimensions are listed in Table I.

TABLE I
THE FEATURE SIZES FOR MULTI-SCALE FUSION.

Feature maps	Size
Low-level feature	$128 \times 28 \times 28$
Mid-level feature	$256 \times 14 \times 14$
High-level feature	$512 \times 7 \times 7$

The MTFE generates $Z_x = \{z_{x,low}, z_{x,mid}, z_{x,high}\}$ representing multiscale features, while the MAFF module produces $F_{out} = \{z_{low}, z_{mid}, z_{high}\}$ indicating fusion features. The CBL module, consisting of 1×1 convolution, BatchNorm, and LeakyRelu, serves to reduce dimensionality, while the Downsample module, consisting of 3×3 convolution, BatchNorm, and LeakyRelu, performs downsampling of the features. The overall calculation is presented as follows.

$$z_{low} = \text{CBL}(\text{ECA}(\text{Concat}(z_{x,low}))) \tag{9}$$

$$z_{mid} = \text{CBL}(\text{ECA}(\text{ECA}(z_{x,mid})) + \text{Down}(z_{low})) \tag{10}$$

$$z_{high} = \text{CBL}(\text{ECA}(\text{Concat}(z_{x,high})) + \text{Down}(z_{mid})), \tag{11}$$

where Down(.) denotes the Downsample module, $x = l, m, s$, and Concat(.) indicates the stitching operation.

D. Multi-scale Transformer Decoder (MTD)

The primary role of the MTD is to decode the encoded feature vectors obtained from various layers, generating three distinct scales of reconstructed images that match the size of the original input image. To enhance the model's overall structure and its ability to reconstruct intricate details, each layer of the decoder is intricately linked with both the Vision Transformer module and the CNN up-sampling module, as previously mentioned. The specific structure is depicted in Fig. 7.

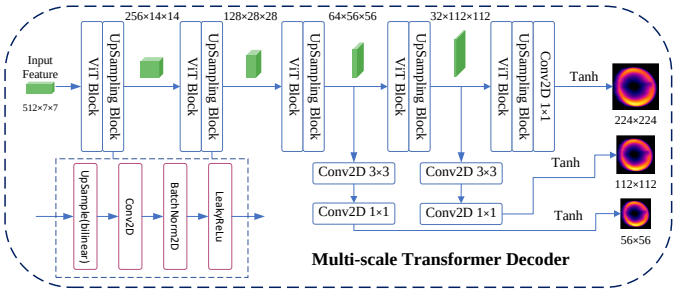


Fig. 7. The structure of the Multi-scale Transformer Decoder (MTD).

The up-sampling module integrates a combination of bilinear interpolation and CNN techniques, departing from the traditional inverse convolutional network approach. This novel strategy is designed to effectively mitigate artifacts encountered during the image reconstruction process. Notably, the final three layers of the MTD are dedicated to a CNN-based dimensionality reduction branch, specifically tailored to reconstruct the target images at varying scales. This design choice enables efficient and accurate reconstruction of images with diverse features and complexities.

E. Discriminator Networks, Bottleneck Structures, and Skip Connections

The discriminator network shares the same basic structure as the multiscale encoder, as depicted in Fig. 3. It includes a multiscale transformer feature extraction module, a multiscale feature fusion module, and a linear layer. The key role of the discriminator network is to form a generative adversarial network with the aforementioned generator module. This collaborative setup aims to elevate the quality and realism of the reconstructed image. Fig. 8 illustrates the bottleneck structure

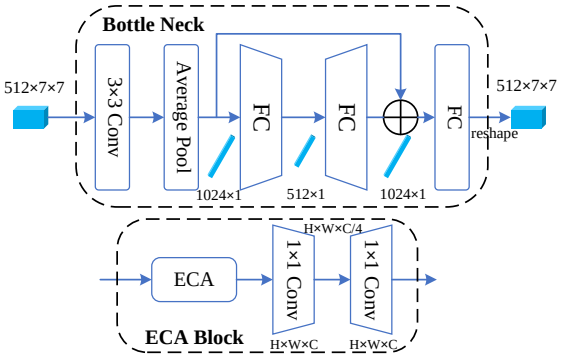


Fig. 8. The architecture of bottleneck and skip connections.

along with the integrated skip connection. The bottleneck structure initially transforms features into a one-dimensional sequence by employing convolution and mean pooling operations. Subsequently, a compact fully connected self-encoder disperses the feature information, aiming to diminish redundancy and subdue atypical features. This process weakens the network's reconstruction ability to reconstruct anomalous features. The skip connection, inspired by U-Net [22], establishes a link between the encoder and decoder, facilitating the preservation of diverse levels of feature information and retaining specific reconstruction details. However, we find that an excessive number of skip connections can prompt the network to reconstruct anomalies without improving detection accuracy. Therefore, as depicted in Fig. 8, we have introduced the ECA module and a simplified bottleneck structure into the skip connections. This novel approach assigns weights to feature channels, suppressing anomalous feature information. The bottleneck structure compresses the feature dimension by a factor of 1/4 via convolution before reinstating the original feature dimension. Notably, skip connections are only established between the lowest-level features.

F. Anomaly Score Module

The anomaly score map S consists of an image-level anomaly score S_i and a feature-level score S_f, $S \in \boldsymbol{R}^{H \times W}$, as depicted in Fig. 3. The image-level anomaly score, denoted as S_i, is computed by applying weights to the Euclidean norm between the multiscale input image $\boldsymbol{I}_{in,x}$ and the multiscale

reconstructed image $\boldsymbol{I}_{out,x}$ at various scales, as indicated by the following equation.

$$S_i = \sum_x \lambda_x \|\boldsymbol{I}_{in,x} - \boldsymbol{I}_{out,x}\|_2, x = l, m, s, \tag{12}$$

where x dignifies three distinct scales, λ_x represents the corresponding weight, satisfying the condition $\sum_x \lambda_x = 1$, and $\|.\|_2$ denotes the Euclidean norm computed across the channels.

The feature-level anomaly score, denoted as S_f, is derived by computing the Euclidean norm of intermediate multiblock features extracted from the encoder and decoder. Specifically, the selected multiblock features, sized of $256 \times 14 \times 14$, $128 \times 28 \times 28$, and $64 \times 56 \times 56$, are uniformly resized to dimensions of 56×56 using linear interpolation. Due to the potential discontinuity existing between different blocks at the feature level, neighboring normal pixels may receive higher anomaly scores. To mitigate this issue, an anomaly score smoothing module is employed. This module executes feature map smoothing through mean pooling utilizing three distinct-sized kernels. Subsequently, the feature-level anomaly scores are obtained by computing the mean after smoothing, as illustrated in Fig. 9.

$$S_f = \text{Up}(\text{Smooth}(\|\boldsymbol{F}_e - \boldsymbol{F}_d\|_2)), S_f \in \boldsymbol{R}^{H \times W} \tag{13}$$

,

$$S = S_i \otimes S_f, \tag{14}$$

where $\boldsymbol{F}_e$ and $\boldsymbol{F}_d$ represent the feature maps after uniform resizing, Smooth(.) means the smoothing process, and Up(.) denotes the upsampling operation.

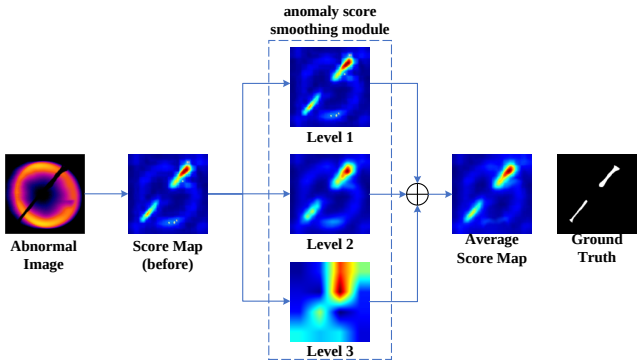


Fig. 9. The architecture of the anomaly score smoothing module.

G. Loss Functions

The core component of our proposed TMIRN is a GAN comprising of a generator (G) and a discriminator (D). Throughout the training process, specific measures are taken to ensure model stability and prevent mode collapse. The discriminator network's loss function (L_D) is optimized using soft labels and gradient penalties, following the principles outlined in the work [35]. Here, $\boldsymbol{I}_{in,x}$ represents the multiscale input image, while $\boldsymbol{I}_{out,x}$ refers to the multiscale reconstructed image. The explicit equations are outlined as follows.

$$L_D = \frac{1}{2}(\|D(\boldsymbol{I}_{in,x}) - 0.9\|_2 + \|D(\boldsymbol{I}_{out,x}) - 0.1\|_2) + \lambda(\|\nabla_{\hat{x}} D(\hat{x})\| - 1)^2, \tag{15}$$

$$\hat{x} = \varepsilon I_{in,x} + (1 - \varepsilon) I_{out,x}, \varepsilon \in [0, 1], \tag{16}$$

where the values 0.9 and 0.1 represents the labels for true and false samples, respectively. The constant coefficient for the gradient penalization term, denoted by λ, is set at a specific value, in this case, as 10. $\nabla_{\hat{x}} D(\hat{x})$ signifies the gradient of the discriminator network to be computed.

The training loss of the generator G combines the content loss with the adversarial loss(L_{adv}). Here, the content loss includes the image reconstruction loss (L_i) and the feature reconstruction loss (L_f). The former incorporates two distinct loss functions: L2 loss and SSIM [21] loss. Meanwhile, the feature reconstruction loss relies on the L2 loss, and can be represented as follows.

$$L_i = \sum_x 1 - \text{SSIM}(\boldsymbol{I}_{in,x}, \boldsymbol{I}_{out,x}) + \sum_x \|\boldsymbol{I}_{in,x} - \boldsymbol{I}_{out,x}\|_2, \tag{17}$$

$$L_f = \|\boldsymbol{F}_e - \boldsymbol{F}_d\|_2, \tag{18}$$

$$L_{adv} = \|D(\boldsymbol{I}_{out,x}) - 1\|_2, \tag{19}$$

$$L_G = \omega_1 L_i + \omega_2 L_f + \omega_3 L_{adv}, \tag{20}$$

where ω_1, ω_2, and ω_3 denote the hyperparameters regulating the effects of the three loss functions, which are empirically set to 0.6, 0.3, and 0.1, respectively.

IV. Experiments and Results

In this section, we will discuss the dataset, the evaluation metrics, and the experimental setup. We will validate the proposed approach by conducting comparison and ablation experiments on the infrared image dataset for aluminum foil sealing. Additionally, we will also demonstrate the generalizability of the proposed approach by utilizing the publicly accessible MVTec AD [36] dataset.

A. Datasets

This study focuses on the identification and localization of defects occurring in aluminum foil seals within the context of the bottling process. The experimental phase of this paper utilizes a proprietary infrared image dataset specifically created for the analysis of aluminum foil sealing. The dataset is acquired from an authentic production setting, with the procedural intricacies visually depicted in Fig. 10.

The infrared imaging process is facilitated by the online thermal imager (MAG32MINI, Shanghai Magnity Technology Co., Ltd., Shanghai), which continuously captures infrared images through photoelectric signals, with a resolution of 384×288. To enhance detection efficiency, the sensing area depicted in Fig. 10 is directly extracted in real-time during the acquisition process by the industrial control all-in-one

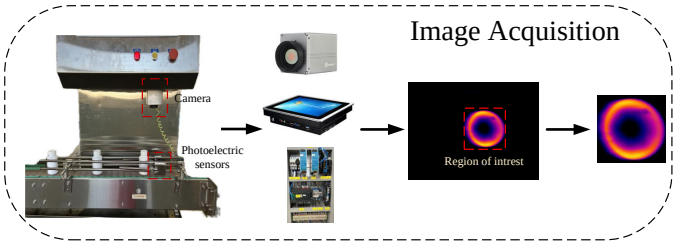


Fig. 10. The data acquisition process.

machine. The resultant dataset includes 400 training images and 150 testing images. All 400 training images are normal, collected in real-time from the production line. In the test set, 85 images exhibit defects and 65 are normal. Among the 85 defective images, only 9 are actual production defects, while the rest are artificially created by damaging aluminum foils and bottle caps. The defects span a spectrum including instances such as absence of aluminum foil, misaligned or loosened caps, notches, fractures, overheating, among others, with a selection of representative images elucidated in Fig. 11.

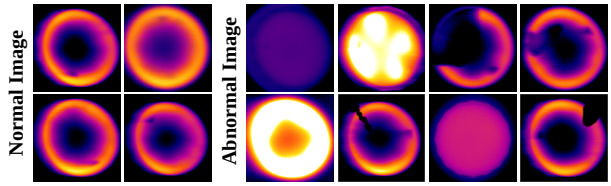


Fig. 11. Selected infrared images for the aluminum foil sealing.

Additionally, this study also assesses the overall generalizability of the proposed network by employing the MVTec AD dataset [36], an established benchmark in the realm of industrial anomaly detection. This dataset is comprised of 5354 high-resolution color images that encompass 5 diverse texture types and 10 varied object types relevant to industrial settings.

Within this dataset, the training subset consists of approximately 200 unlabeled normal images, whereas the test subset comprises a mixture of both normal and anomaly images, featuring a total of 70 distinct industrial defects, including instances such as staining, missing, and broken, among others. These images have pixel sizes ranging between 700 and 1024. A selection of these images is provided in Fig. 12.

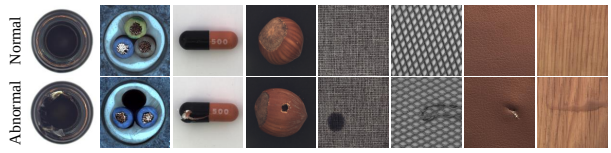


Fig. 12. Selected images from the MVTec AD dataset [36].

B. Evaluation Metrics

This paper adopts the area under the receiver operating characteristic curve (AUROC) and the area under the per-region-overlap curve (AUPRO) as primary evaluation metrics. For anomaly detection, we employ image-level AUROC (I-AUROC). For anomaly localization, we consider pixel-level AUROC (P-AUROC) and AUPRO. The AUROC comprises the True Positive Rate (TPR) and the False Positive Rate (FPR), wherein its assessment of anomaly localization tends to favor the detection of larger-area anomalies. Conversely, The AUPRO assigns equal importance to anomalous regions of varying sizes, thereby offering a more accurate assessment of anomaly localization performance.

C. Implement Details

The proposed TMIRN model is implemented on a Python 3.8-based deep learning framework using PyTorch 2.0.0. All experiments are conducted on a Windows system equipped with a i9-12900H 2.50 GHz processor and an NVIDIA GeForce GTX 3060 GPU graphics card. During the training process, the model is trained solely on normal samples and undergoes sample augmentation by randomly flipping and rotating the image by 10 degrees, with an input image size of $224 \times 224 \times 3$. The model is trained for 200 epochs using the Adam optimizer with a learning rate of 2e-4 and a Beta parameter of (0.5,0.999). The detailed performance metrics and resource requirements of the model are shown in Table II. During testing, a mixture of normal and anomalous samples is used as input for evaluation.

TABLE II
PERFORMANCE METRICS AND RESOURCE REQUIREMENTS OF THE TMIRN MODEL.

Metric	Model Size (MB)	Maximum GPU Memory Usage (GB)	Number of Parameters (M)	Computational Cost (GMac)
Result	301.88	4.97	84.9	7.91

D. Comparative Results

In order to evaluate the efficacy of the proposed framework for detecting defects of infrared images, this study conducts a comparative analysis against the state-of-the-art methodologies. Given the limited availability of infrared images for defect detection in existing research, several unsupervised learning methods suitable for anomaly detection and localization on the MVTec AD dataset have been selected for comparison. The selected methods include eight baseline approaches: AnoGAN [17], f-AnoGAN [16], AE-SSIM [12], Skip-GANomaly [15], DRAEM [33], InTra [30], Patchcore [26], and RD++ [27] Methods.

The comparison experiments are conducted using three evaluation metrics: I-AUROC, P-AUROC, and AUPRO, on the infrared image dataset for aluminum foil sealing. Each model has been trained five times, and the results are averaged for robustness. A summary of the comparative results is presented in Table III.

We can find that the proposed model in this paper demonstrates a significant performance enhancement when applied to the infrared image dataset for aluminum foil sealing, outperforming the other eight baseline methods. In terms of anomaly detection, our model surpasses the DRAEM network by 0.85% and achieves a remarkable accuracy of 99.25%. Concerning the anomaly localization task, the pixel-level AUROC metric of our TMIRN achieves the highest segmentation

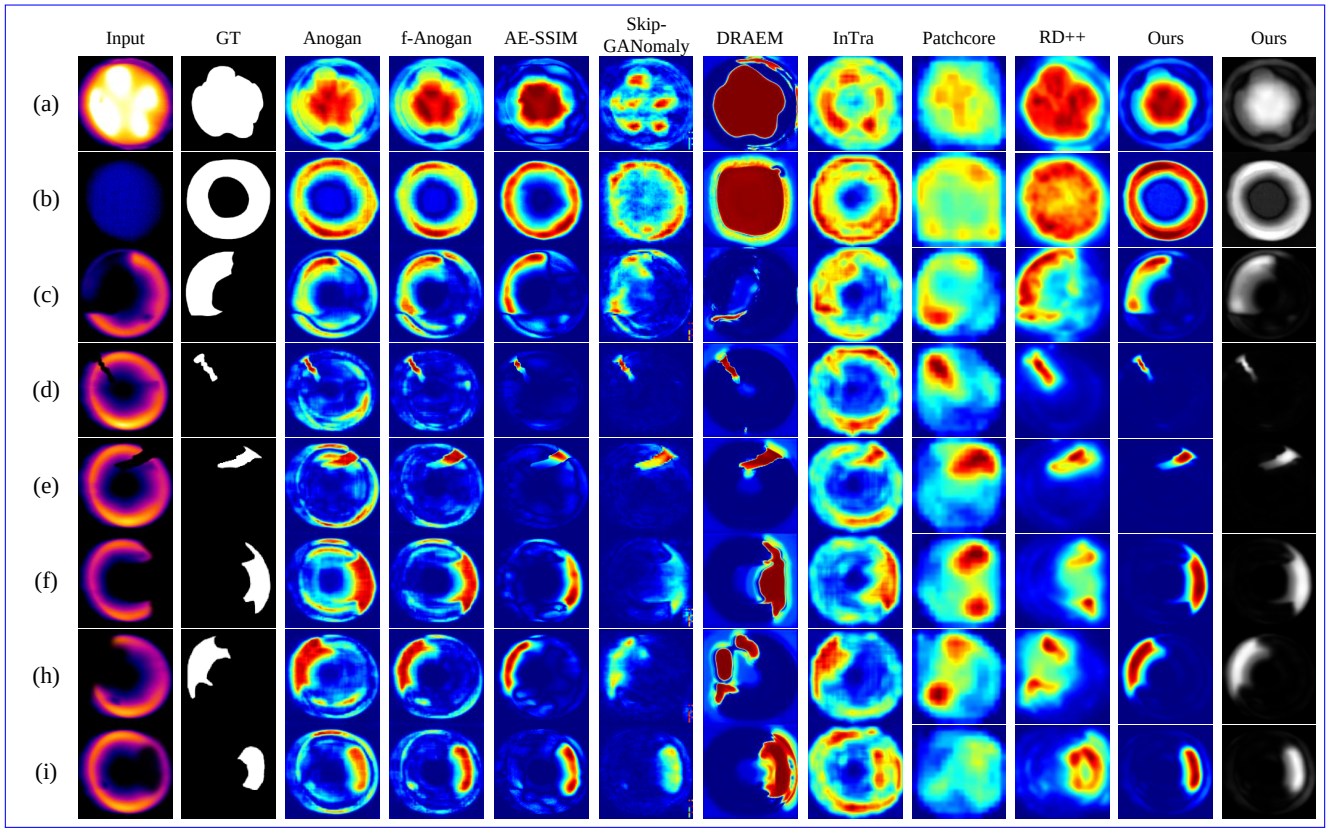


Fig. 13. Evaluation of our proposed method on the infrared image dataset in comparison with AnoGAN [17], f-AnoGAN [16], AE-SSIM [12], Skip-GANomaly [15], DRAEM [33], InTra [30], Patchcore [26], and RD++ [27].

TABLE III
RESULTS OF ANOMALY DETECTION AND LOCALIZATION ON THE INFRARED IMAGE DATASET.

Metrics / Model	Detection	Localization	
	I-AUROC(%)	P-AUROC(%)	AUPRO(%)
AnoGAN	81.21±0.07	78.37±0.05	59.31±0.05
f-AnoGAN	92.92±0.06	92.02±0.04	74.16±0.04
AE-SSIM	97.90±0.05	91.83±0.06	77.90±0.06
Skip-GANomaly	79.33±0.06	92.52±0.07	72.41±0.04
DRAEM	98.40±0.07	72.17±0.05	71.29±0.07
InTra	88.72±0.05	82.63±0.07	74.54±0.09
Patchcore	91.83±0.03	88.71±0.04	75.85±0.05
RD++	98.91±0.04	92.86±0.06	80.11±0.05
TMIRN(ours)	**99.25**±0.06	**96.41**±0.06	**86.61**±0.04

accuracy of 96.41%. Moreover, our model exhibits superior performance compared to AE-SSIM, which holds the second-highest accuracy, by 8.71% on the more strict AUPRO metric. Moreover, Fig. 13 depicts the detailed results of anomaly detection and localization for both the methods proposed in this paper and the eight baseline methods. The figure includes anomaly images displaying eight distinct scale defects, ground truth images, anomaly segmentation heat maps obtained from different networks, and anomaly score maps generated by the networks proposed in this study. The results can clearly indicate that our TMIRN exhibits superior accuracy in accurately localizing the anomaly region for defects of varying scales and shapes.

Based on the analysis of images (c), (e), and (f), it is clear that image-level reconstruction networks such as AnoGAN and f-AnoGAN tend to introduce a significant number of interfering pixels. To address this issue, the method proposed in this paper combines image-level anomaly scores with feature-level anomaly scores to minimize the interference caused by these pixels. Skip-GANomaly performs well in reconstructing medium- and large-scale anomalous features, mainly due to the excess of redundant information conveyed through multiple skip connections. However, DRAEM struggles with segmenting large-scale features, particularly in types (b), (c), and (h). InTra, which relies on patch-based anomaly localization, shows noticeable boundary effects in its anomaly score map. PatchCore, meanwhile, fails to accurately identify large-scale defects that closely resemble the background, primarily due to the patch-based approach's lack of global semantic information. Similarly, RD++ faces challenges in fully recognizing large-scale defects. Therefore, we can claim that our method outperforms the eight baseline methods when applied to the infrared image dataset for aluminum foil sealing.

E. Ablation Study

In this subsection, we have conducted a series of ablation experiments probing various facets of the proposed the TMIRN framework, demonstrating its enhanced effectiveness. We also provide a concise description of the methodology and analytical procedures employed in this study.

1) The Number of Skip Connections

In this paper, we implement feature information transfer by establishing skip connections between the encoder and decoder components at matching hierarchical levels. This augmentation strategy supplies the model with finer-grained features for image reconstruction. However, it is important to note that an overabundance of transferred information may lead to a decrease in detection accuracy due to the false reconstruction of the anomalous features. Therefore, we empirically examine the impact of the quantity of skip connections on the training effectiveness of the model.

Fig. 14 shows the comparison of the detection accuracy of the models with varying number of skip connections. The incremental integration of skip connections, originating from zero and progressing from the highest-level encoder feature, is visually depicted in the accuracy curve. Notably, the utilization of a singular skip connection demonstrates peak performance across all three evaluation metrics. As the number of skip connections increases, the model exhibits heightened proficiency in reconstructing anomalous features, significantly impacting the accuracy of image anomaly segmentation. Conversely, the absence of skip connections impedes the reconstruction of intricate image details, consequently leading to a notable decline in image-level detection accuracy. Further insight into the corresponding image reconstruction outcomes can be found in Fig. 15.

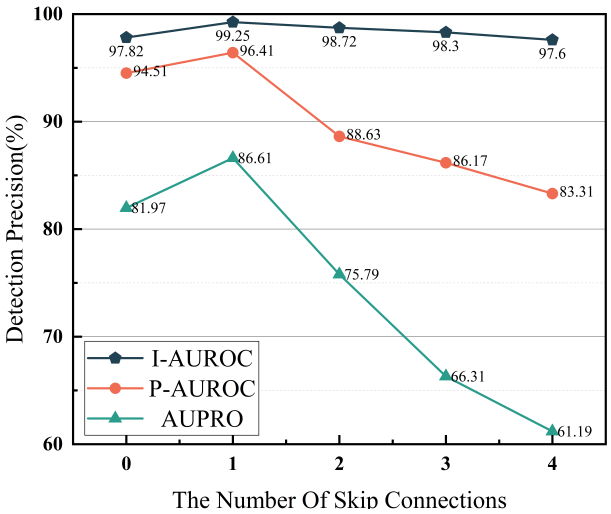


Fig. 14. The impact of the quantity of skip connections on the model's performance. The number of skip connections refers to the total layers starting from the deep features of the encoder that utilize skip connections.

2) Multi-scale Input and Feature Fusion Module

In this series of experiments, we investigate the impact of the multi-scale input and multi-scale feature fusion module on the efficacy of the proposed TMIRN. The experiments assess the training outcomes of single-scale input (without MAFF), multi-scale input (without MAFF), and multi-scale input (with MAFF), respectively. The single-scale encoder network consists of five layers, utilizing the residual block and Vision Transformer block as the base modules, with feature output sizes of 112, 56, 28, 14, and 7. A simple direct fusion mode is employed in multi-scale input (without MAFF). As shown in Fig. 3, features extracted from three different sizes of images with a 7×7 output are concatenated and fused using a simple convolution. Analysis of the results are presented in Fig. 16 and Table IV. As shown in Table IV,

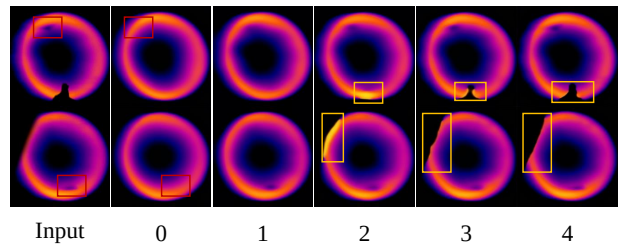


Fig. 15. Alterations in image reconstruction are observable across varying quantities of skip connections. The red boxed area corresponds to the intricate reconstruction of image details, while the yellow boxed region represents the reconstruction of anomalous regions.

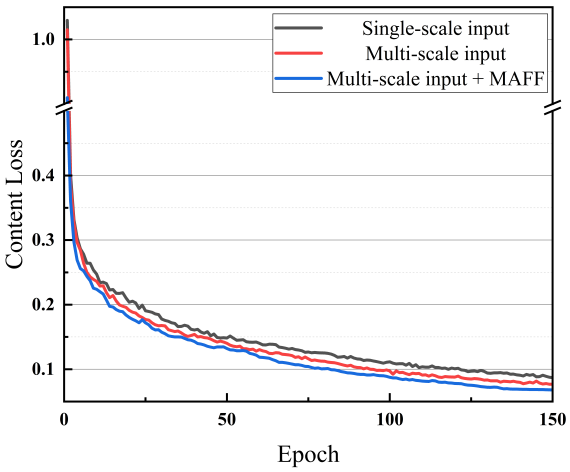


Fig. 16. Content loss of different scale inputs and feature fusion in the training process.

TABLE IV
DEFECT DETECTION AND LOCALIZATION RESULTS WITH DIFFERENT SCALE INPUTS AND FEATURE FUSION.

Methods \ Metrics	Detection	Localization		Runtime (ms)
	I-AUROC(%)	P-AUROC(%)	AUPRO(%)	
Single-scale input	98.12	92.31	81.22	**21.3**
Multi-scale input (without MAFF)	99.21	95.17	84.96	27.9
Multi-scale input (with MAFF)	**99.25**	**96.41**	**86.61**	29.1

multi-scale inputs significantly enhance detection accuracy by enabling the model to focus on features at different scales. The MAFF module further improves performance by effectively integrating essential feature information, resulting in optimal detection outcomes. The comparison between single-scale input, multi-scale input (without MAFF), and multi-scale input (with MAFF) highlights the advantages of the MAFF module. Specifically, the I-AUROC improves by 1.13%, and the P-AUROC improves by 4.10%. These results demonstrate the MAFF module's effectiveness in handling multi-scale features, capturing complex characteristics, and enhancing the model's performance in detection and localization tasks. The addition of the MAFF module results in only a 1ms increase in runtime,

TABLE V
DEFECT DETECTION AND LOCALIZATION RESULTS FOR DIFFERENT FEATURE EXTRACTION MODULES.

Residual Convolution	Vision Transformer	Detection	Localization		Runtime (ms)
		I-AUROC(%)	P-AUROC(%)	AUPRO(%)	
		97.44	93.28	82.63	19.4
✓		98.25	94.11	83.74	20.0
	✓	99.01	96.07	85.98	27.7
✓	✓	99.25	96.41	86.61	29.1

while the overall increase for multi-scale inputs compared to single-scale inputs is approximately 7ms. The runtime for single-scale input is 21.3ms, whereas for multi-scale input (without MAFF) and multi-scale input (with MAFF), the runtimes are 27.9ms and 29.1ms, respectively. This slight increase in time is acceptable in industrial applications given the significant performance gains.

Overall, the combination of multi-scale inputs and the MAFF module not only enhances detection accuracy but also improves feature extraction efficiency, making it highly effective for practical applications.

3) Feature Extraction Module in MTFE

To enhance the model's ability to capture global contextual information and accurately identify critical fine-grained defects, this paper integrates residual convolution blocks with the Vision Transformer within the MTFE module. This specific architecture aims to leverage the strengths of both convolutional and transformer frameworks. To validate the effectiveness of this module, we have also conducted detailed experimental analyses.

As shown in Table V, the full MTFE module, which includes both residual convolution and Vision Transformer modules, achieves the highest performance, with an I-AUROC of 99.25% and an AUPRO of 86.61%. The residual convolution module improves the performance to an I-AUROC of 98.25% and an AUPRO of 83.74%, while the Vision Transformer alone reaches an I-AUROC of 99.01% and an AUPRO of 85.98%. The improvements indicate that residual convolutions capture fine-grained details, while the Vision Transformer enhances global context, and their combination effectively boosts defect detection and localization. Despite a slight increase in runtime, the accuracy gains justify this inclusion.

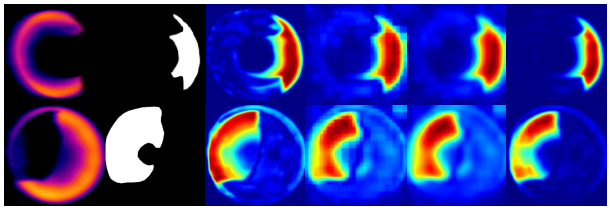
4) Abnormal Score Module

The model presented in this paper conducts defect detection and segmentation via anomaly scores, employing a module that combines image-level anomaly scores and feature-level anomaly scores while applying feature-level anomaly score smoothing. In this experiment, we validate the effectiveness of the anomaly score module by comparing the detection accuracy of various anomaly scores. The results, detailed in Table VI, demonstrate that the combination of image-level anomaly scores and pixel-level anomaly scores outperforms the individual use of these scores in terms of accuracy. Deriving image-level anomaly scores directly from pixel space differentials restricts their capacity to filter out noise produced during the reconstruction process. Conversely, feature-level anomaly scores prioritize abnormal regions but sacrifice detailed information.

Moreover, the incorporation of the smoothing module amplifies the overall impact of feature-level anomaly scores. This optimization effect is notably illustrated in Fig. 17, where the anomaly score module described this paper effectively mitigates erroneous pixels resulting from image detail reconstruction, concurrently enhancing the clarity of feature-level anomaly scores at the fuzzy boundaries.

TABLE VI
DETECTION AND LOCALIZATION RESULTS OF DIFFERENT ANOMALY SCORING METHODS.

Methods \ Metrics	Detection	Localization	
	I-AUROC(%)	P-AUROC(%)	AUPRO(%)
Image-level	98.51	94.63	81.76
Feature-level	98.01	94.90	81.97
Feature-level + smoothing	99.03	95.50	83.95
Image-level + smoothed feature-level	**99.25**	**96.41**	**86.61**



(a) (b) (c) (d) (e) (f)

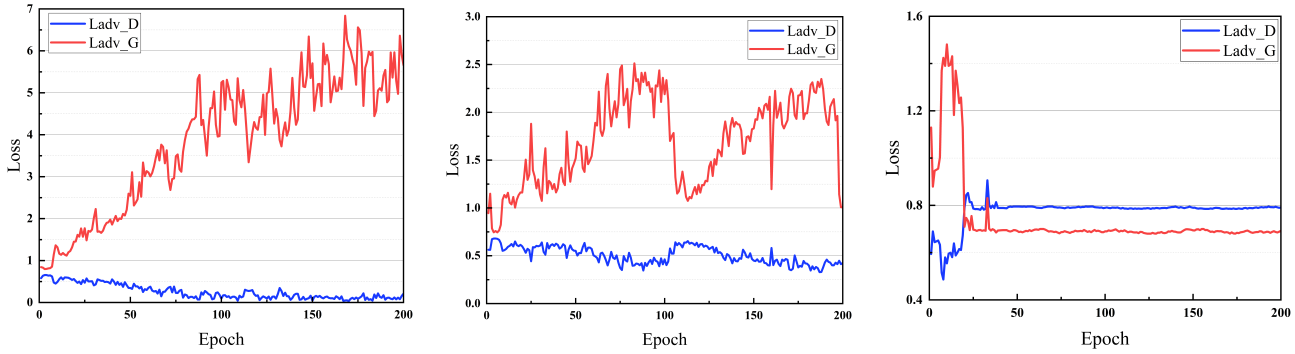
Fig. 17. Plot of anomaly scores for different methods: (a) the anomalous image, (b) the ground truth, (c) the image-level, (d) the feature-level, (e) the smoothed feature-level, and (f) the combination of the image-level and the smoothed feature-level.

5) Soft Labelling and Gradient Penalties

The adversarial loss function of the discriminator in the generative adversarial network incorporates soft labels and gradient penalties. This subsection focuses on empirically verifying their impact on the model's stability during adversarial training. As depicted in Fig. 18, the adversarial loss curves of the generator and discriminator reveal significant insights. When employing conventional training labels (0, 1), the discriminator's loss rapidly converges to 0, leading to an imbalance in generator training and a common issue known as model collapse with GAN training. Replacing these labels with soft labels (0.1, 0.9) prevents the discriminator from quickly converging to critical levels, thus effectively mitigating pattern collapse. However, the training process remains unstable in terms of adversarial loss, making it challenging to achieve convergence. The introduction of the gradient penalty introduces a gradual convergence of the training loss following a transient disturbance, ultimately leading to the attainment of the Nash equilibrium state. This observation highlights the effectiveness of the gradient penalty in fostering stability and facilitating convergence in the adversarial training process.

F. Generalizability Testing on The MVTec AD

In this experiment, we assess the generalizability of the anomaly detection and localization performance of the proposed TMIRN model using the MVTec AD [36], a comprehensive industrial dataset comprising 5 texture types and 10 object



(a) (0, 1) labeling. (b) (0.1, 0.9) Soft labeling. (c) Soft labeling and gradient penalties.

Fig. 18. Adversarial losses of different methods.

categories. The primary objective is to validate the model's general applicability as proposed in this paper. We compare the results with three other image reconstruction-based networks, AE-SSIM [12], GANomaly [18], and AnoViT [37]. The results in Table VII clearly demonstrate that our TMIRN outperforms the other three models in terms of overall anomaly detection and localization. The TMIRN model consistently outperforms across most categories, confirming its strong generalizability. However, it's worth noting that our TMIRN exhibits slightly reduced effectiveness in detecting anomalies within complex texture categories. Fig. 19 visually illustrates the results of the TMIRN's anomaly localization for all categories in the MVTec AD dataset, showcasing its adaptability in detecting various defect types of varying scales and complexities.

TABLE VII
ANOMALY DETECTION AND LOCALIZATION RESULTS ON THE MVTEC AD [36] IN COMPARISON WITH AE-SSIM [12], GANOMALY [18] AND ANOVIT [37], IN TERMS OF THE EVALUATION METRICS OF IMAGE-LEVEL AUROC (%) / PIXEL-LEVEL AUROC (%).

Class/Model	AE-SSIM	GANomaly	AnoViT	TMIRN
Carpet	**87.01**/64.73	84.23/55.02	50.01/65.07	86.12/**89.87**
Grid	94.02/84.91	74.32/80.07	52.05/83.00	**97.71/94.85**
Leather	78.07/56.15	79.21/77.07	85.05/89.02	**95.62/97.04**
Tile	59.09/47.55	79.54/69.05	**89.07**/57.02	87.89/**86.51**
Wood	73.02/60.34	65.33/**91.01**	95.05/85.07	**96.47**/90.91
Bottle	93.06/83.47	89.21/82.02	83.02/86.06	**99.41/97.04**
Cable	82.02/47.81	73.25/83.07	74.01/89.06	**92.99/94.10**
Capsule	**94.05**/86.04	70.81/72.07	73.06/91.05	93.21/**96.64**
Hazelnut	97.01/91.67	79.44/86.05	88.01/94.05	**98.07/99.09**
Metal nut	89.08/60.35	74.57/69.01	86.09/88.05	**92.12/94.50**
Pill	**91.05**/83.01	75.77/76.06	72.01/86.03	**91.05/95.63**
Screw	96.05/88.79	69.98/72.04	**100.00**/92.07	99.06/**97.82**
Toothbrush	92.01/78.45	70.01/82.03	74.07/90.08	**99.20/98.91**
Transistor	90.02/72.50	74.63/79.04	83.07/80.01	**92.72/94.10**
Zipper	88.07/66.58	83.47/84.06	73.05/76.05	**91.31/97.07**
Average	86.93/69.41	76.24/77.12	78.56/83.41	**94.20/94.94**

V. CONCLUSIONS AND FUTURE WORK

In the domain of industrial infrared image anomaly detection, the challenges such as non-uniform backgrounds, noise interference, and the absence of detailed semantic feature information impede the efficacy of image reconstruction-based models. Similarly, anomaly detection methods relying on feature embedding may encounter issues related to category bias, while the dynamic contours present in infrared images pose difficulties in ensuring consistent object localization. Addressing these challenges, this paper introduces a novel approach called Transformer-based multi-scale image reconstruction network (TMIRN), designed specifically for the infrared images of aluminum foil sealing. The proposed TMIRN innovatively combines elements from generative adversarial networks, incorporating both CNN and Transformer architectures to leverage both global and local semantic information effectively. To improve defect localization across varying scales, the model adopts a multi-scale framework that merges image-level multi-scale anomaly scores with feature-level scores. This integration aims to diminish the influence of background noise prevalent in infrared images.

Experimental results demonstrate the superior accuracy of our TMIRN in both detection and localization on a proprietary dataset. The validation of TMIRN's generalization on the MVTec AD dataset and its superior performance compared to other image reconstruction networks highlight its potential impact on the broader industrial anomaly detection field. Its ability to achieve real-time detection, with a remarkable speed of 0.029 seconds per image, is particularly noteworthy as it meets the critical real-time requirements essential for industrial applications. This capability positions TMIRN as a promising solution for scenarios where immediate anomaly detection is crucial, such as in manufacturing processes or quality control environments.

However, despite these strengths, our TMIRN still faces challenges in handling complex textures like Carpet and Tile within the MVTec AD dataset. This limitation is attributed to the trade-off between suppressing defect reconstruction and accurately capturing intricate texture details. Improving TMIRN's ability to reconstruct heterogeneous texture features effectively will not only enhance its performance on specific datasets but also contribute to advancing anomaly detection techniques for diverse industrial applications.

In future studies, we will explore potential distributions for normal samples and investigate feature clustering methodologies to augment the TMIRN's capacity. The objective is

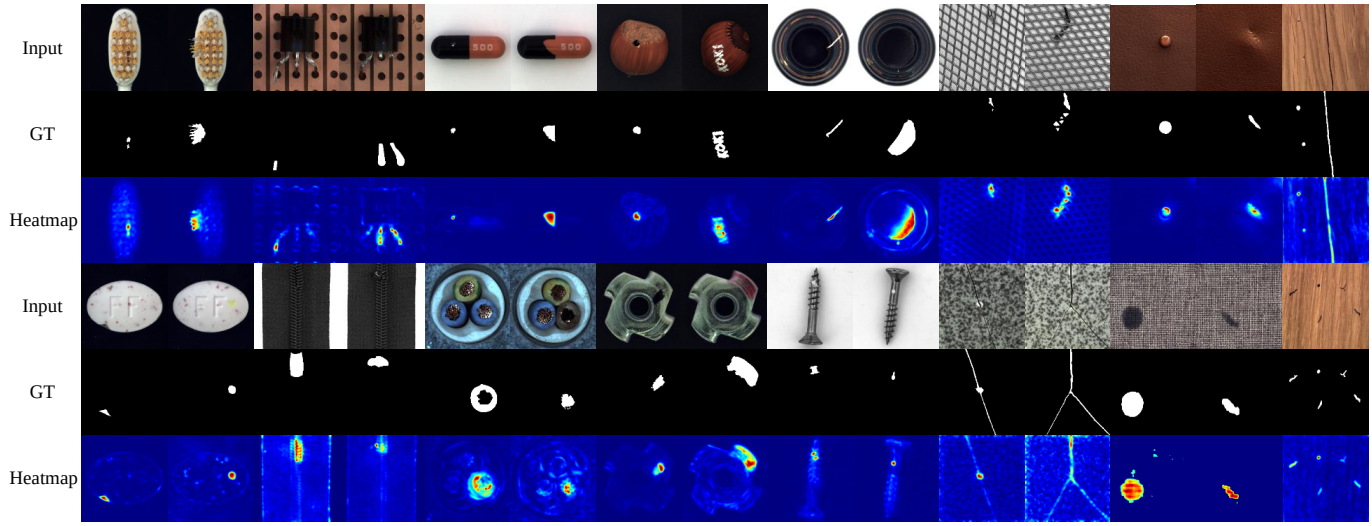


Fig. 19. Heatmap of the anomaly scores of the proposed TMIRN model on the MVTec AD [36] dataset.

to enhance its ability to reconstruct intricate textures while maintaining anomaly detection accuracy, thereby broadening its applicability across various industrial settings. Moreover, in practical industrial implementations, the neural network-based framework demonstrates efficacy but demands considerable computational resources and memory allocation. Our future work will also focus on model compression techniques. The objective is to sustain operational efficiency while adhering to existing resource constraints and computational prerequisites.

REFERENCES

[1] S. Cruz, A. Paulino, J. Duraes, and M. Mendes, "Real-time quality control of heat sealed bottles using thermal images and artificial neural network," *Journal of Imaging*, vol. 7, no. 2, p. 24, 2021.

[2] Y. Gao, X. Li, X. V. Wang, L. Wang, and L. Gao, "A review on recent advances in vision-based defect recognition towards industrial intelligence," *Journal of Manufacturing Systems*, vol. 62, pp. 753–766, 2022.

[3] O. Janssens, M. Loccufier, and S. Van Hoecke, "Thermal imaging and vibration-based multisensor fault detection for rotating machinery," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 1, pp. 434–444, 2018.

[4] Y. He, B. Deng, H. Wang, L. Cheng, K. Zhou, S. Cai, and F. Ciampa, "Infrared machine vision and infrared thermography with deep learning: A review," *Infrared Physics & Technology*, vol. 116, p. 103754, 2021.

[5] N. Rajic, "Principal component thermography for flaw contrast enhancement and flaw depth characterisation in composite structures," *Composite Structures*, vol. 58, no. 4, pp. 521–528, 2002.

[6] Z. Yan, C.-Y. Chen, L. Luo, and Y. Yao, "Stable principal component pursuit-based thermographic data analysis for defect detection in polymer composites," *Journal of Process Control*, vol. 49, pp. 36–44, 2017.

[7] J. Ahmed, B. Gao, and W. L. Woo, "Wavelet-integrated alternating sparse dictionary matrix decomposition in thermal imaging cfrp defect detection," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 7, pp. 4033–4043, 2018.

[8] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[9] J. Luo, Z. Yang, S. Li, and Y. Wu, "Fpcb surface defect detection: A decoupled two-stage object detection framework," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–11, 2021.

[10] C. Li, H. Yan, X. Qian, S. Zhu, P. Zhu, C. Liao, H. Tian, X. Li, X. Wang, and X. Li, "A domain adaptation yolov5 model for industrial defect inspection," *Measurement*, vol. 213, p. 112725, 2023.

[11] Y. He, K. Song, Q. Meng, and Y. Yan, "An end-to-end steel surface defect detection approach via fusing multiple hierarchical features," *IEEE Transactions on Instrumentation and Measurement*, vol. 69, no. 4, pp. 1493–1504, 2019.

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

[13] T. Niu, B. Li, W. Li, Y. Qiu, and S. Niu, "Positive-sample-based surface defect detection using memory-augmented adversarial autoencoders," *IEEE/ASME Transactions on Mechatronics*, vol. 27, no. 1, pp. 46–57, 2021.

[14] 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 Transactions on Automation Science and Engineering*, vol. 16, no. 3, pp. 1450–1467, 2019.

[15] S. Akçay, A. Atapour-Abarghouei, and T. P. Breckon, "Skip-ganomaly: Skip connected and adversarially trained encoder-decoder anomaly detection," in *International Joint Conference on Neural Networks*, pp. 1–8, IEEE, 2019.

[16] T. Schlegl, P. Seeböck, S. M. Waldstein, G. Langs, and U. Schmidt-Erfurth, "f-anogan: Fast unsupervised anomaly detection with generative adversarial networks," *Medical Image Analysis*, vol. 54, pp. 30–44, 2019.

[17] 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 *International Conference on Information Processing in Medical Imaging*, pp. 146–157, Springer, 2017.

[18] S. Akcay, A. Atapour-Abarghouei, and T. P. Breckon, "Ganomaly: Semi-supervised anomaly detection via adversarial training," in *Asian Conference on Computer Vision*, pp. 622–637, Springer, 2019.

[19] M. Niu, Y. Wang, K. Song, Q. Wang, Y. Zhao, and Y. Yan, "An adaptive pyramid graph and variation residual-based anomaly detection network for rail surface defects," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–13, 2021.

[20] Y. Shi, J. Yang, and Z. Qi, "Unsupervised anomaly segmentation via deep feature reconstruction," *Neurocomputing*, vol. 424, pp. 9–22, 2021.

[21] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[22] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention*, pp. 234–241, Springer, 2015.

[23] V. Zavrtanik, M. Kristan, and D. Skočaj, "Reconstruction by inpainting for visual anomaly detection," *Pattern Recognition*, vol. 112, p. 107706, 2021.

[24] T. Defard, A. Setkov, A. Loesch, and R. Audigier, "Padim: A patch distribution modeling framework for anomaly detection and localization," in *International Conference on Pattern Recognition*, pp. 475–489, Springer, 2021.

[25] L. Chen, Z. You, N. Zhang, J. Xi, and X. Le, "Utrad: Anomaly detection and localization with u-transformer," *Neural Networks*, vol. 147, pp. 53–62, 2022.

[26] K. Roth, L. Pemula, J. Zepeda, B. Schölkopf, T. Brox, and P. Gehler, "Towards total recall in industrial anomaly detection," in *Proceedings of*

the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14318–14328, 2022.
[27] T. D. Tien, A. T. Nguyen, N. H. Tran, T. D. Huy, S. Duong, C. D. T. Nguyen, and S. Q. Truong, "Revisiting reverse distillation for anomaly detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24511–24520, 2023.
[28] Z. Liu, Y. Zhou, Y. Xu, and Z. Wang, "Simplenet: A simple network for image anomaly detection and localization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20402–20411, 2023.
[29] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021.
[30] J. Pirnay and K. Chai, "Inpainting transformer for anomaly detection," in *International Conference on Image Analysis and Processing*, pp. 394–406, Springer, 2022.
[31] X. Tao, C. Adak, P.-J. Chun, S. Yan, and H. Liu, "Vitalnet: Anomaly on industrial textured surfaces with hybrid transformer," *IEEE Transactions on Instrumentation and Measurement*, vol. 72, pp. 1–13, 2023.
[32] A. Mariani and G. Malucelli, "Insights into induction heating processes for polymeric materials: An overview of the mechanisms and current applications," *Energies*, vol. 16, no. 11, p. 4535, 2023.
[33] V. Zavrtanik, M. Kristan, and D. Skočaj, "Draem-a discriminatively trained reconstruction embedding for surface anomaly detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8330–8339, 2021.
[34] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pvt v2: Improved baselines with pyramid vision transformer," *Computational Visual Media*, vol. 8, no. 3, pp. 415–424, 2022.
[35] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *International Conference on Machine Learning*, pp. 214–223, PMLR, 2017.
[36] P. Bergmann, M. Fauser, D. Sattlegger, and C. Steger, "Mvtec ad–a comprehensive real-world dataset for unsupervised anomaly detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9592–9600, 2019.
[37] Y. Lee and P. Kang, "Anovit: Unsupervised anomaly detection and localization with vision transformer-based encoder-decoder," *IEEE Access*, vol. 10, pp. 46717–46724, 2022.

Dr. Changyun Wei received the PhD degree in Artificial Intelligence from the Delft University of Technology, Netherlands, in 2015. Currently, he is an Associate Professor with the College of Mechanical and Electrical Engineering, Hohai University, China. He has published more than 30 papers in the fields of computer vision, artificial intelligence and robotics. His research interests include robotics, multi-agent systems, intelligent systems and computer vision.

Hui Han received the B.S. degree in Mechanical Engineering from Nanjing Institute of Technology, Nanjing, China, in 2022. He is currently pursuing the M.S. degree in the College of Mechanical and Electrical Engineering, Hohai University, China. His research interests are industrial machine vision, deep learning and defect detection.

Zhichao Wu received the B.S. degree in Mechanical Engineering from Nanjing Institute of Technology, Nanjing, China, in 2023. He is currently pursuing the M.S. degree in the College of Mechanical and Electrical Engineering, Hohai University, China. His research interests are industrial vision inspection, deep learning and image processing .

Yu Xia received the B.S. degree in Mechanical Engineering from Hohai University of Wentian College, Ma'anshan, China, in 2020. He is currently pursuing the M.S. degree in the College of Mechanical and Electrical Engineering, Hohai University, China. His research interests are deep learning and industrial vision inspection.

Dr. Ze Ji received the PhD. degree from Cardiff University, Cardiff, UK, in 2007. He is a Senior Lecturer (Associate Professor) with the School of Engineering, Cardiff University, and the recipient of the Royal Academy of Engineering Industrial Fellowship. Prior to his current position, he was working in industry (Dyson, Lenovo, etc) on autonomous robotics. His research interests are cross-disciplinary, including autonomous robot navigation, robot manipulation, robot learning, computer vision.