SELF-SUPERVISED DEFECT DETECTION AND LOCALIZATION BASED ON HEATMAP PSEUDO ANOMALIES (HPA)

WENLEI LIU and ZIYU ZHU

Anomaly detection is widely used in manufacturing and medical imaging. We propose a self-supervised defect detection method based on multi-scale feature fusion, which can effectively improve the detection and localization accuracy. The method of pseudo-defect construction was used to enhance the training data. To make the pseudo-defects more realistic, the extreme point of feature heatmap was used as the anchor point of the defect area, and the defect image was fused with the original image to construct the pseudo-defect. A multi-scale feature fusion network was proposed that utilizes the self-attention mechanism and the interaction between multi-scale features to extract semantic features containing rich contextual information to improve detection and localization accuracy further. The proposed method achieved competitive experimental results on both the MVTec AD and Chest X-ray datasets. Compared with other pseudo-defect simulation methods, the heatmap-based pseudo-defect construction method improves by at least 2%. It achieves comparable results with other state-of-the-art defect detection methods.

Key words: self-supervised, defect detection, heatmap pseudo anomalies, multi-scale feature fusion

Introduction

With the rapid development of computer vision, deep learning methods have been widely used for defect detection in industrial production. With the continuous improvement of automation technologies, the defect rate of products has been getting lower and lower, making the collection of sufficient defective samples more and more difficult. Samples with sporadic defects cannot even be collected. Insufficient defect data limits the application of data-driven methods. At the same time, labeling defect data requires a great deal of manpower and material resources, with some defects less recognizable, and some labeling work requires professional knowledge background. For the above reasons, supervised methods requiring a large amount of labeled information are unsuitable for defect detection tasks. In contrast, self-supervised methods that require little or no labeled data are better for defect detection.

Due to the small volume of defective samples, data augmentation methods are usually used to expand the labeled data. Studies have shown that data augmentation strategies that simulate natural defects can effectively improve the accuracy of image anomaly detection. CutPaste Li et al. (2021) is a commonly used data augmentation method that cuts a part of an image and randomly pastes it

---

Wenlei Liu is a postdoctoral fellow at the Institute of Artificial Intelligence, Department of Computer Science, Tsinghua University. Ziyu Zhu is a doctoral student at the Department of Artificial Intelligence, Department of Computer Science, Tsinghua University.

AI Magazine. 00(000): 1–12; doi:10.1002/aaai.0001

into another image to anomaly defects. However, the defects have evident discontinuous edge traces, which can easily cause overfitting during the training process. FPI Tan et al. (2020) is a pseudo-defect construction method for synthetic anomalies, which extracts the same patch area from two independent samples, uses the interpolation between the two patches to replace the original patch area, and obtains the interpolation factor, patch size, and patch location by random sampling. FPI is suitable for detection of defects of texture and defects in medical imaging. NSA Schlüter et al. (2021) uses Poisson image editing to seamlessly blend blocks of different sizes and scales in the image, making the synthetic anomalies more like irregular anomalies of natural images.

Self-supervised methods do not require prior knowledge and only need standard training data. Anomalies are inferred by the difference between the test data and learned regular features. Therefore, self-supervised methods are widely used in few-shot defect detection. Current self-supervised methods can be divided into two broad categories: reconstruction-based and pre-trained-based methods. In the absence of anomaly samples and pre-trained models, reconstruction-based models, such as the Variational Automatic Encoder (VAE) Liu et al. (2020) and Generative Adversarial Networks (GAN) Perera, Nallapati, and Xiang (2019), are widely used. Generative models consist of encoders and decoders. An auto-encoder that reconstructs standard data is first trained with anomalous-free data. During inference, anomalies are detected and localized by comparing pixel-level differences between input and generated images. Self-supervised methods based on pre-trained usually need to use the pre-trained model to extract the feature vector of the input image. The similarity between the features of the abnormal image and those of the pre-trained feature vector will be calculated, and the abnormal score will be obtained, based on which it will be judged whether an abnormality exists. Self-supervised methods based on pre-trained have poor generalization ability and network interpretability, and thus are often used in conjunction with other self-supervised methods.

In this paper, the method of pseudo-defect construction was adopted to increase the number of training samples, and a self-supervised method based on semantic segmentation was used for defect detection and localization. CutPaste, FPI, and NSA methods usually use randomly selected regions to construct defective samples. Still, random methods have a problem: unrealistic defects, such as faults and hollows in the sample image, or a defect patch in the blank part of the image, often occur. These issues do not affect the presence of image defects, but the constructed pseudo-defects are unrealistic. In this paper, the point with a higher intensity value in the feature heatmap was used as the anchor point to select the patch area so that the constructed defects are more realistic. Figure 1 shows the comparison of different pseudo-defect construction methods. Self-supervised defect detection methods based on pixel-level reconstruction errors and probability density anomalies cannot capture high-level semantic information. Therefore, a U-net structure for semantic segmentation and fuses multi-scale features was adopted to improve self-supervised defect detection methods. The main innovations of this paper are as follows:

- We propose to use the point with a higher intensity value in the feature heatmap of the image as an anchor to select the candidate region and then fuse the target image and the patch, thus constructing reasonable pseudo-defects.
- A self-supervised network structure based on multi-scale feature fusion was proposed, and the attention mechanism was used to increase the expressiveness of features. Semantic feature information can be extracted by this network, which is used for defect detection and segmentation.
- This method achieved good experimental results on both the MVTec AD and Chest X-ray datasets.

Related Work

There are many methods for defect detection, which can be divided into three categories: self-supervised and unsupervised methods, generative model-based methods, and semantic segmentation methods.

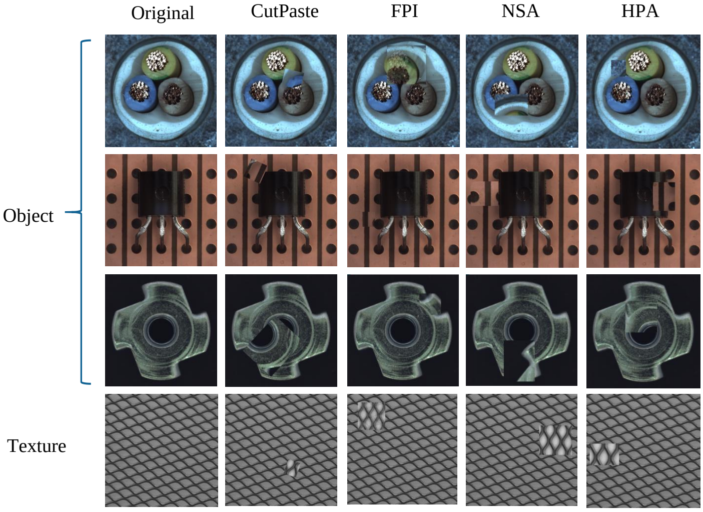

Figure 1. Comparison of different pseudo-defect synthesis methods.

Self-supervised and unsupervised defect detection methods

Knowledge distillation is a commonly used self-supervised defect detection method in which some knowledge of a teacher network is transferred to a student network. Usually, a pre-trained strong network is selected as the teacher network, which can fully extract the features of the image Perera, Nallapati, and Xiang (2019); Ristea et al. (2021). The student network has the same structure as the teacher network. The teacher network guides the student network to learn the distribution of non-anomalous images to retain critical features. The multi-scale hierarchical feature matching method can enable the student network to effectively learn multi-level mixed knowledge from the feature pyramid to detect anomalies of various scales Bergmann et al. (2020); Yamada and Hotta (2021). Deng *et al.* Deng and Li (2022) proposed a paradigm of ”reverse distillation”, consisting of a teacher encoder and a student decoder. Instead of directly accepting the original image, the student network takes the output features of the teacher network as input to recover the teacher’s multi-scale representation of the goal. Another commonly used self-supervised processing method is using one-class data to learn the representation of a self-supervised model and then using the model to detect the defects of different objects Sohn et al. (2020); Yi and Yoon (2020); Tan et al. (2020).

Unsupervised defect detection usually assumes that a model trained only on standard samples exhibits “maladaptation” to abnormal data Bergmann et al. (2019). Fastflow Yu et al. (2021) is a probability distribution estimator with a deep feature extractor (*ResNet*, Visual Transformer) for unsupervised anomaly detection and localization. Fastflow learns to convert visual features into tractable probability distributions and obtains probabilities of anomalies during the inference stage. Zheng *et al.* Zheng et al. (2021) proposed an unsupervised anomaly detection and localization method aligned from coarse to delicate in normal images. The coarse alignment stage normalizes the positions of image feature pixels, and the delicate alignment stage calculates the similarity of corresponding position features. Staged unsupervised methods are widely used in defect detection and usually include a feature extraction stage and an anomaly detection scoring stage Song et al. (2021); Wan et al. (2021); Liu, Zhuang, and Lu (2021); Cohen and Hoshen (2020).

Defect detection method based on generative model

The generative model is significant in the field of deep learning. Standard generative models for defect detection include VAE Liu et al. (2020), GAN Akcay, Atapour-Abarghouei, and Breckon (2018), and the flow model.

Defect detection with generative models typically involves two steps: first, training an auto-encoder on anomaly-free data, then during inference, minimizing the $L2$ distance between the output representation and the reference point and calculating pixel-level differences to obtain precise anomaly locations Massoli et al. (2021); Yang, Shi, and Qi (2020); Roth et al. (2021). Zavrtanik *et al.* Zavrtanik, Kristan, and Skočaj (2021) proposed a Discriminatively trained Reconstructed Anomaly Embedding Model (DRAEM), which learns the joint representation of anomalous images and reconstructions without anomalies and simultaneously learns the decision boundary between normal and abnormal data. This method can directly locate anomalies without requiring sophisticated post-processing. Pirnay *et al.* Pirnay and Chai (2022) proposed that anomaly detection can be studied as a patch repair problem and proposed a self-attention mechanism-based detection method that utilizes the generated parts to patch defect locations. Liang *et al.* Liang et al. (2022) decoupled the input image into components with different frequencies and modelled the reconstruction process as a parallel restored combination of full-frequency images. Since there are significant differences between the frequency distributions of normal and abnormal images, this method can effectively determine defect locations.

Defect detection method based on semantic segmentation

There are many kinds of defect detection methods based on semantic segmentation. Pre-trained models are widely used in semantic segmentation, and applying pre-trained CNNs or visual transformers to segmentation can effectively improve the defect detection and localization accuracy Defard et al. (2021); Fort, Ren, and Lakshminarayanan (2021). The matching probability of multi-scale features can improve the accuracy of semantic segmentation, which has strong robustness to noise Sohn et al. (2020); Kwon et al. (2020); Rudolph, Wandt, and Rosenhahn (2021). Generally, global features can be used to detect whether an object has defects, while local features can be sued to determine the specific location of anomalies. Therefore, the combination of global and local features is widely used in defect detection Kamoona et al. (2021); Rudolph et al. (2022). Rippel *et al.* Rippel, Mertens, and Merhof (2021) used multivariate Gaussian distribution to represent features at different scales with a normal model and used the variance of standard data to distinguish normal data from abnormal data. Using contrastive learning in visual recognition tasks can also improve detection accuracy Peng et al. (2022). Reiss *et al.* Reiss and Hoshen (2021) normalized the extreme points of the features obtained by contrastive learning and scaled them into a unit sphere, effectively improving defect detection accuracy by using constraints.

Method

Heatmap Pseudo Anomalies (HPA)

In defect detection, most data are from normal samples, only a small amount of data contain defects, and there are no defective data in the training samples. Therefore, this paper adopts the method of pseudo-defect construction to increase the number of pseudo-labels in self-supervised training.

The pseudo-defect construction method based on the heatmap used in this paper is mainly divided into four steps as follows:

- 1) Using the feature heatmap of the training data, an extreme point on the heatmap of an image is obtained as the anchor point of the defect area;
- 2) An intensity extreme point is used as the anchor point of the candidate box, and the Gamma function is used to adjust the height and width of the candidate box;
- 3) A scaling coefficient is randomly selected from a standard normal distribution, and the scale of the candidate box is adjusted;
- 4) Taking an extreme point on the heatmap of the target image as the anchor point, the area matching the candidate box of the source image is selected for fusion, thus obtaining the image of the pseudo-defect.

The method to obtain the feature heatmap in this paper is using $ResNet50$ to extract the features of the source image, conveying the feature distribution F, and then using the following formula to normalize the features:

$$f_{norm} = \frac{f - f_{min}}{f_{max} - f_{min}}, \tag{1}$$

where f_{min} is the minimum eigenvalue and f_{max} is the maximum eigenvalue.

The normalized features are sorted, and the image position corresponding to the most significant feature is selected as the anchor point of the candidate box. Multiple candidate anchor points are usually chosen according to a particular feature threshold to select a suitable candidate box.

The defects of objects are usually minor and have specific local characteristics, so we use the *Gamma* function to select the height h and width w of the candidate box. The formula is as follows:

$$h = G_a \cdot H, w = G_a \cdot W, \tag{2}$$

where G_a is the Gamma function, H is the height of the source image, and W is the width of the source image.

Because the shapes and sizes of defects in different objects are different, the size of the candidate box is adjusted by the scale factor s, which is selected from a standard normal distribution with a mean of 1 and a variance of 0.5. The height h_s and width w_s of the adjusted candidate box are as follows: $h_s = s \cdot h, w_s = s \cdot w$.

In the target image, an extreme point on the heatmap is used to select the fusion region that matches the candidate box of the source image. During matching, if the target area is not chosen correctly, an appropriate point should be re-selected from the candidate extreme points on the heatmap.

Finally, the images of pseudo-defects are re-annotated. The defect location of the target image is annotated, which can be conveniently determined through semantic segmentation during the self-supervised training process. At the same time, the entire pseudo-defect image is annotated to facilitate the identification of defect images during the training process.

Model Framework

The self-supervised framework proposed in this paper is shown in Figure 2. The encoder part is equivalent to the teacher network in knowledge distillation, and the decoder part is equivalent to the student network. The defect detection is performed by comparing the features of the two regions. During training, the pseudo-defect image is taken as input and pre-processed into 224×224 pixels. The training goal is to identify the defective images in the input sequence and locate the defect position.

The encoder part uses a pre-trained $ResNet$, and the pre-trained network parameters are the model parameters for $ResNet$ to perform classification training on the ImageNet dataset. The encoder and decoder have a similar network structure, making it convenient to calculate feature similarity. The strides of convolutional layer and pooling layer are taken as 2, and each block extracts features of different levels. The high-resolution features of the bottom levels include color, texture, and edge information, while the low-resolution features of the top levels contain rich contextual information. Fusing features at different levels can complement information and improve recognition accuracy. The decoder mainly restores the underlying features from the top-level features, expands the dimensional difference between different layers by up-sampling between different layers, and uses bilinear interpolation to smooth the up-sampled image.

The similarity score represents the similarity of the features acquired by the teacher and student networks, which is usually used to judge whether there is a defect. The higher the similarity score, the higher the possibility of a defect. Assuming that F_{ei} is the encoder feature of the i-th layer and that F_{di} is the decoder feature of the i-th layer, the similarity score S_i of the i-th layer is:

$$S_i = \frac{{F_{ei}}^T \cdot F_{di}}{\|F_{ei}\| \cdot \|F_{di}\|}, \tag{3}$$

The loss function of the training process according to the similarity score of each layer can be calculated as:

$$L = \frac{1}{N}\sum_{i=1}^{N} L_i = \frac{1}{N}\sum_{i=1}^{N}(1 - S_i), \tag{4}$$

where S_i is the similarity score of the i-th layer, L_i is the loss function of the i-th layer, and N is the number of layers of encoder and decoder. In this paper, $N = 3$.

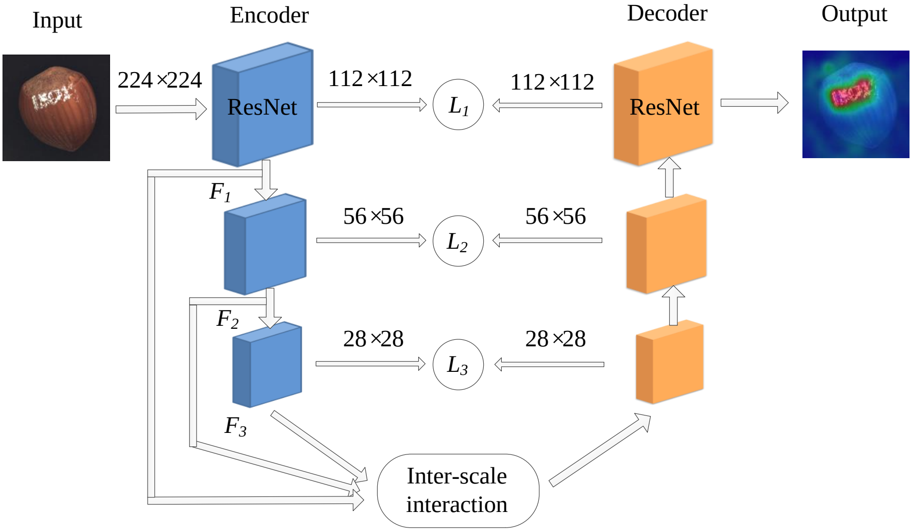

Figure 2. The self-supervised framework consists of three parts: encoder, inter-scale interaction, and decoder.

Multi-scale feature fusion network

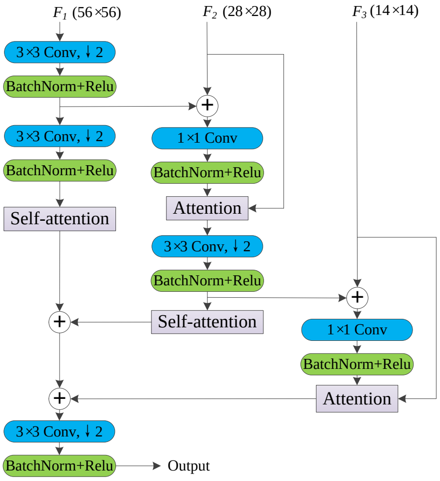

Figure 3. Multi-scale feature fusion network structure.

The multi-scale feature fusion network used in this paper is shown in Figure 3, in which the 3×3 convolution realizes the information interaction of different layers and learns local features. Because the number of channels in different layers is different, to achieve a compelling fusion of features of different layers, 1×1 convolution transforms the number of channels to align two tensors with the same size but a different number of channels. Because inputs of different scales are very likely to lead to various weight updates, the direction of the optimizer's minimization is unbalanced, which leads to a disproportionate shape of the loss function, thus reducing the convergence rate of the training process. Therefore, the data must be normalized during the training process to improve the speed and accuracy of training. The nonlinear activation function used in this paper is the *Relu* activation function.

The main idea of multi-scale feature fusion is as follows: taking the feature F_i of the current layer as the main feature, taking the feature $F_{i-1} (i \geq 2)$ of the upper layer as the auxiliary feature, and making the features of different scales interact through the fusion method. The fused feature is $F_i^{'} = Fusion(F_{i-1}, F_i)$, and finally the interaction feature is applied to the current layer by using the attention mechanism. After that, the features can be obtained as follows:

$$\begin{aligned} F_{ai} &= \mathrm{softmax}(\frac{Q \cdot K^T}{\sqrt{M_i}})V \\ &= \mathrm{softmax}(\frac{F_i^{'} \cdot F_i^T}{\sqrt{M_i}})F_i, \end{aligned} \tag{5}$$

where Q, K and V refer to the Query, Key and Value of the attention mechanism, respectively. M_i is the dimension of the i-th layer feature.

The self-attention mechanism can capture rich and detailed information in the image, suppress useless information, highlight essential features, and obtain the correlation between data, thereby improving the accuracy of defect recognition. This paper uses a self-attention module in each layer to capture features, and its calculation is similar to Eq. 5, where $Q = K = V = F_i$.

Finally, the outputs of all the processed layers are fused, and the final multi-scale fusion features are obtained through the 3 × 3 convolution layer, the normalization layer, and the *Relu* function.

Experiment

Experimental setup

Dataset: This paper mainly uses the MVTec AD and Chest X-ray datasets to conduct experiments to verify and compare the proposed method with the state-of-the-art defect detection methods. The MVTec AD dataset imitates the industrial production scene and provides pixel-level annotations for anomalous regions. The dataset contains 5 textures and 10 objects in different fields and has rich anomaly types. Because there are usually a small number of anomalies, the training set contains only standard samples, and the test set includes both normal and abnormal examples. The Chest X-ray dataset contains about 110,000 chest radiographs from different patients, and the problem of anomaly detection in chest radiographs can be viewed as a general image classification problem. The dataset has the following characteristics: many chest radiographs are visually very similar; multiple labels may be required to label a chest radiograph to represent different diseases; the data is unbalanced, and there are far more normal instances than abnormal instances. Therefore, there are only standard samples in the training process and anomalous samples in the test set. The training sets of the above two datasets only contain standard samples, so self-supervised or unsupervised methods are needed to learn the feature representation of normal samples and detect abnormal samples.

Parameter setting: The pixel of the dataset image used in this paper is $700 \times 700 - 1024 \times 1024$. To facilitate the extraction of features by the proposed method, it's necessary to resize the image to 256×256 pixels and use image enhancement methods such as random rotation and flipping. Finally, the image is randomly cropped to 224×224. This paper's encoder and decoder parts are mainly composed of $ResNet34$ modules with fewer model parameters and can extract higher-level abstract features.

Evaluation criteria: The detection problem can be divided into two steps: anomalous image classification and anomalous region segmentation. The abnormal image classification only outputs whether the image is an abnormal sample. The irregular area segmentation needs to judge whether there is an abnormality and find the abnormal area. Defect detection is usually evaluated using the "Receiver Operating Characteristic" (ROC), and the area enclosed by the ROC curve and the coordinate axis is called $AUROC$. The reason for using $AUROC$ as the evaluation index is that the index is not sensitive to the threshold or the percentage of abnormality. The larger the $AUROC$, the higher the possibility that there is an abnormality. To verify the repeatability and robustness of the method proposed in this paper, the defect detection and localization experiments were performed 5 times, and then the average value was taken as the final experimental result.

MVTec AD dataset experiment

To verify the effectiveness of the self-supervised defect detection method based on heatmap pseudo anomalies (HPA) proposed in this paper, it is compared with several state-of-the-art self-supervised and unsupervised detection methods, among which CutPaste Li et al. (2021) and NSA Schlüter et al. (2021) are methods based on pseudo-defect; RD4AD Deng and Li (2022) is a method based on knowledge distillation; DRAEM Zavrtanik, Kristan, and Skočaj (2021) is a method based on generative models; and CS-Flow Rudolph et al. (2022) and FastFlow Yu et al. (2021) are methods based on flow models.

Defect detection experiment. The experimental results of defect detection are shown in Table 1. The results show that for most

Table 1. **Defect detection comparison results on MVTec dataset**

Category	CutPaste	NSA	DRAEM	RD4AD	CS-Flow	FastFlow	HPA
carpet	**100**$_{\pm0.0}$	95.6$_{\pm0.6}$	97.0	98.9	98.7	**100.0**	99.5$_{\pm0.1}$
grid	99.1$_{\pm0.0}$	99.9$_{\pm0.1}$	99.9	**100.0**	99.6	99.7	**100**$_{\pm0.0}$
leather	**100**$_{\pm0.0}$	99.9$_{\pm0.1}$	**100.0**	**100.0**	**100.0**	**100.0**	**100**$_{\pm0.0}$
tile	99.8$_{\pm0.2}$	100$_{\pm0.0}$	99.6	99.3	99.9	**100.0**	98.8$_{\pm0.2}$
wood	99.8$_{\pm0.0}$	97.5$_{\pm1.5}$	99.1	99.2	99.1	**100.0**	99.6$_{\pm0.1}$
bottle	**100**$_{\pm0.0}$	97.7$_{\pm0.3}$	99.2	**100.0**	**100.0**	**100.0**	**100**$_{\pm0.0}$
cable	96.2$_{\pm0.3}$	94.5$_{\pm1.0}$	91.8	95.0	97.5	**100.0**	99.0$_{\pm0.1}$
capsule	95.4$_{\pm0.1}$	95.2$_{\pm1.7}$	98.5	96.3	97.7	**100.0**	97.9$_{\pm0.2}$
hazelnut	99.9$_{\pm0.0}$	94.7$_{\pm1.1}$	100.0	99.9	99.9	**100.0**	**100**$_{\pm0.0}$
metalnut	98.6$_{\pm0.0}$	98.7$_{\pm0.7}$	98.7	**100.0**	99.2	**100.0**	**100**$_{\pm0.0}$
pill	93.3$_{\pm0.2}$	99.2$_{\pm0.6}$	98.9	96.6	96.8	**99.4**	98.3$_{\pm0.2}$
screw	96.6$_{\pm0.2}$	90.2$_{\pm0.6}$	93.9	97.0	91.9	**97.8**	97.6$_{\pm0.1}$
toothbrush	90.7$_{\pm0.1}$	**100**$_{\pm0.0}$	**100.0**	99.5	99.6	94.4	**100**$_{\pm0.0}$
transistor	97.5$_{\pm0.2}$	95.1$_{\pm0.2}$	93.1	96.7	95.2	**99.8**	99.7$_{\pm0.0}$
zipper	99.9$_{\pm0.1}$	99.8$_{\pm0.1}$	**100.0**	98.5	98.5	99.5	99.3$_{\pm0.1}$
Average	97.0$_{\pm0.0}$	97.2$_{\pm0.3}$	98.0	98.5	98.3	**99.4**	99.2$_{\pm0.1}$

Table 2. **Defect localization comparison results on MVTec dataset**

Category	CutPaste	NSA	DRAEM	RD4AD	CS-Flow	FastFlow	HPA
carpet	98.3$_{\pm0.0}$	95.5$_{\pm2.3}$	95.5	98.9	99.3	**99.4**	98.3$_{\pm0.0}$
grid	97.5$_{\pm0.1}$	99.2$_{\pm0.1}$	**99.7**	99.3	99.0	98.3	99.0$_{\pm0.0}$
leather	99.5$_{\pm0.0}$	99.5$_{\pm0.1}$	98.6	99.4	**99.6**	99.5	99.3$_{\pm0.1}$
tile	90.5$_{\pm0.2}$	**99.3**$_{\pm0.0}$	99.2	95.6	98.0	96.3	96.0$_{\pm0.4}$
wood	95.5$_{\pm0.1}$	90.7$_{\pm1.9}$	96.4	95.3	96.6	**97.0**	96.3$_{\pm0.3}$
bottle	97.6$_{\pm0.1}$	98.3$_{\pm0.1}$	**99.1**	98.7	98.9	97.7	98.5$_{\pm0.2}$
cable	90.0$_{\pm0.2}$	96.0$_{\pm1.4}$	94.7	97.4	97.6	**98.4**	97.1$_{\pm0.1}$
capsule	97.4$_{\pm0.1}$	97.6$_{\pm0.9}$	94.3	98.7	99.0	**99.1**	98.1$_{\pm0.1}$
hazelnut	97.3$_{\pm0.0}$	97.6$_{\pm0.6}$	**99.7**	98.9	98.9	99.1	98.1$_{\pm0.1}$
metalnut	93.1$_{\pm0.4}$	98.4$_{\pm0.2}$	**99.5**	97.3	98.5	98.5	97.2$_{\pm0.1}$
pill	95.7$_{\pm0.1}$	98.5$_{\pm0.3}$	97.6	98.2	98.9	**99.2**	98.3$_{\pm0.0}$
screw	96.7$_{\pm0.1}$	96.5$_{\pm0.1}$	97.6	**99.6**	98.9	99.4	99.2$_{\pm0.0}$
toothbrush	98.1$_{\pm0.0}$	94.9$_{\pm0.7}$	98.1	**99.1**	98.9	98.9	98.8$_{\pm0.1}$
transistor	93.0$_{\pm0.2}$	88.0$_{\pm1.8}$	90.9	92.5	98.0	97.3	**98.6**$_{\pm0.2}$
zipper	**99.3**$_{\pm0.0}$	94.2$_{\pm0.4}$	98.8	98.2	99.0	98.7	98.3$_{\pm0.1}$
Average	96.0$_{\pm0.1}$	96.3$_{\pm0.4}$	97.3	97.8	**98.6**	98.5	98.1$_{\pm0.2}$

objects, a detection accuracy of 99% - 100% can be achieved with the method proposed in this paper, which is comparable to the detection accuracy of the most advanced detection methods, whether it is a texture defect or object defect. Among the textures and objects in the MVTec AD, screws, capsules, and pills have the worst detection accuracies, below 99%. They have common characteristics: these objects are relatively small and usually have no obvious texture information, making it difficult to detect defects through texture information. These tiny defects are very likely to be missed in inspection.

Compared with CutPaste and NSA that are based on pseudo-defect construction, the accuracy of the proposed method in this paper is significantly improved. The method of locating the defect center by heatmap can generate higher-quality defective samples than the original pseudo-defect construction method. In terms of defect detection accuracy, the proposed method is also better than the three methods of DRAEM, RD4AD, and CS-Flow. The reason is that the proposed method obtains the relationship between feature contexts through the multi-scale feature fusion method. The features of higher level and broader perception fields are accepted simultaneously, which can effectively improve detection accuracy. There is a small gap in defect detection accuracy between the proposed method and the FastFlow method based on the flow model. The proposed method has better robustness than the FastFlow method.

Defect location experiment. The experimental results of defect location are shown in Table 2. The results show that the proposed method can locate defects more accurately and stably than other pseudo-defect construction methods. Pseudo-defect construction is a very effective data enhancement method, which can effectively reduce the problem of overfitting training data and obtain higher localization accuracy. Because there are only standard samples in the training data, if an inappropriate data augmentation method is used, the generalization ability of the learned defect segmentation model will be poor. In this paper, the defect segmentation method combining U-net and multi-scale feature fusion network was used, which can effectively fuse the features of different layers of networks, expand the segmentation detection area, and realize the interaction between various layer features, thus learning and locating defect more accurately.

Comparative experiment of different pseudo-defect construction methods

The effects of different pseudo-defect construction methods on defect detection were compared through experiment. In the experiment, all methods used the self-supervised training network based on the multi-scale feature fusion proposed in this paper. The W/O group is a control group in which all samples in the training data are standard samples. Other groups, including CutPaste, FPI Tan et al. (2020), NSA, and HPA mentioned in this paper, add pseudo-defect samples to the training data.

The comparison results of different pseudo-defect construction methods are shown in Table 3, and the heatmap of the defect location is shown in Figure 4. The experimental results show that when there are only standard samples in the training data, the defect detection accuracy is low, and effective detection cannot be carried out, but it has little effect on defect localization. This is because the U-net network used in this paper can effectively learn the texture features of different layers and then segment abnormal data according to the learned features to achieve accurate defect location. When pseudo-defect samples are added to the training data, the accuracies of detection and localization significantly increase to more than 97%. The more realistic the constructed pseudo-defects, the more significant their effects on defect detection. It's because adding near-realistic defects to the training data is equivalent to adding labeled data to the training data. Usually, the more the labeled data, the better the training effect of the model. Using pseudo-defects can achieve the result of supervised training while reducing the cost and workload of data labeling. Therefore, pseudo-defects are crucial in the training of self-supervised models.

Chest X-ray dataset ablation experiment

This paper used the Chest X-ray dataset to compare different pseudo-defect construction methods. The localization and detection results are shown in Table 4. The feature heatmap of defect localization is shown in Figure 5, where various diseases are labeled on the left side of the graph. The network used in this experiment is a self-supervised training network based on multi-scale feature fusion, and the input is training data constructed with different pseudo-defect methods. Chest X-ray radiographs were divided into male and female in the experiment, and their detection and localization results were respectively verified.

By comparing the experimental results of Chest X-ray and MVTec AD, it can be found that the detection and localization results of MVTec AD are better than those of Chest X-ray. The reason is that the standard samples in the MVTec AD dataset are similar to each other, and the defects of the objects are more noticeable, the defect outline and texture information can better serve the purpose of detection and localization. However, there are significant differences between standard samples in the Chest X-ray dataset containing normal samples due to the different body structures of patients. At the same time, the tube or cardiac equipment on the patient may interfere with the detection. Moreover, the radiographs are relatively blurry, their textures are not clear enough, and the difference between the sick and the normal radiographs is also tiny, which increases the difficulty of detection and localization. Figure 5 shows

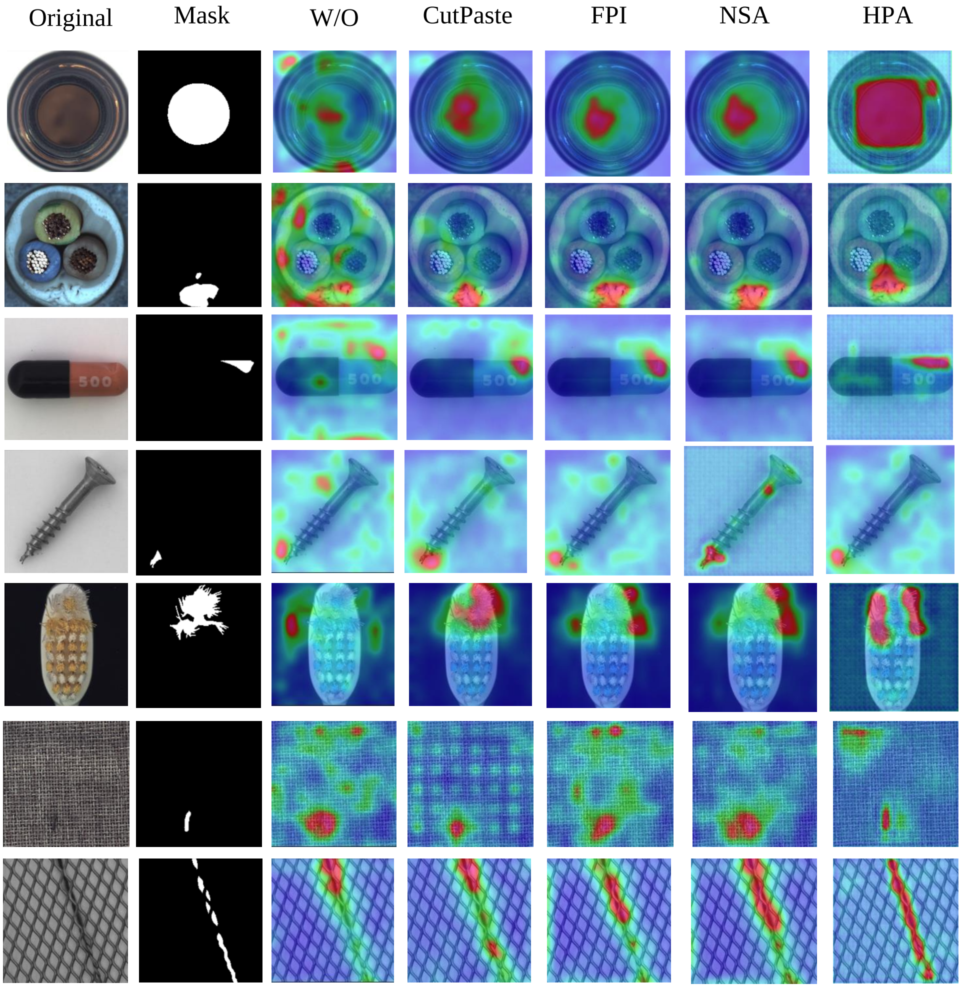

Figure 4. Feature heatmap of MVTec AD defect localization.

Table 3. Comparison of experimental results of different pseudo-defect construction methods of MVTec AD. The experimental results are the average detection and location results of different objects.

	W/O	CutPaste	FPI	NSA	HPA
Detection	$87.3_{\pm 0.1}$	$97.6_{\pm 0.2}$	$98.8_{\pm 0.1}$	$98.1_{\pm 0.1}$	$\mathbf{99.2}_{\pm 0.2}$
Localization	$91.0_{\pm 0.2}$	$97.0_{\pm 0.2}$	$97.3_{\pm 0.1}$	$97.6_{\pm 0.1}$	$\mathbf{98.1}_{\pm 0.2}$

that in defect localization, besides the proposed method, the method based on FPI defect construction can also achieve good segmentation results. The reason is that the X-ray images are relatively blurry, and FPI is a construction method that uses interpolation to synthesize anomalies, resulting in blurred pseudo-defects, which is very suitable for defect detection in medical imaging.

Conclusion

We proposed a self-supervised defect detection method based on heatmap pseudo-defect construction to improve the detection and localization accuracy. The traditional pseudo-defect construction method was enhanced to construct more realistic pseudo-defects by using an extreme point on

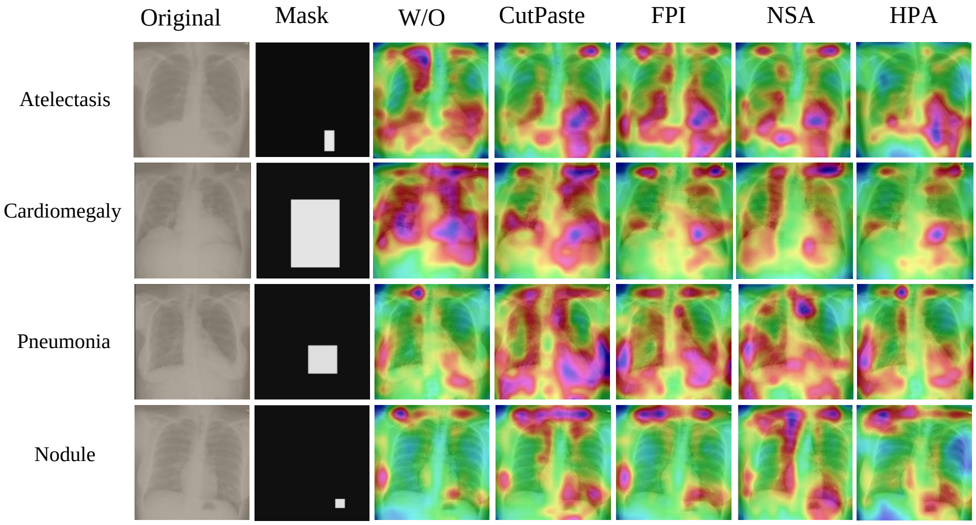

Figure 5. Feature heatmap of Chest X-ray defect location.

Table 4. Comparative experimental results on Chest X-ray dataset with different pseudo-defect construction methods

		W/O	CutPaste	FPI	NSA	HPA
Detection	Male	$88.2_{\pm 1.8}$	$90.2_{\pm 1.5}$	$93.2_{\pm 1.1}$	$95.2_{\pm 1.1}$	$\mathbf{97.6}_{\pm 1.1}$
	Female	$88.7_{\pm 1.7}$	$90.5_{\pm 1.6}$	$93.8_{\pm 0.8}$	$95.4_{\pm 0.9}$	$\mathbf{97.5}_{\pm 1.2}$
Localization	Male	$83.3_{\pm 2.1}$	$89.6_{\pm 1.2}$	$92.4_{\pm 1.1}$	$91.8_{\pm 0.8}$	$\mathbf{94.6}_{\pm 1.2}$
	Female	$83.2_{\pm 1.8}$	$89.7_{\pm 1.2}$	$92.1_{\pm 1.0}$	$92.0_{\pm 0.3}$	$\mathbf{94.7}_{\pm 1.0}$

the feature heatmap as the anchor point of the defect area and fusing the defect image into the original image. Experimental results showed that pseudo-defects could improve defect detection and localization accuracy and have good generalization ability. The more realistic the pseudo-defect, the higher the defect detection and localization accuracy. A self-supervised network structure of multi-scale feature fusion was designed to obtain the semantic features of the contextual information of different layers, enhancing its feature expression ability. The proposed method achieved good results on both the MVTec AD and Chest X-ray datasets. There is still much room for improvement for the application of defect detection in medical imaging. In the future, combining pseudo-defects with generative models can be considered to improve the applicability and accuracy of self-supervised defect detection methods.

References

Akcay, S.; Atapour-Abarghouei, A.; and Breckon, T. P. 2018. Ganomaly: Semi-supervised anomaly detection via adversarial training. In *Asian conference on computer vision*, 622–637. Springer.

Bergmann, P.; Fauser, M.; Sattlegger, D.; and Steger, C. 2019. MVTec AD–A comprehensive real-world dataset for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9592–9600.

Bergmann, P.; Fauser, M.; Sattlegger, D.; and Steger, C. 2020. Uninformed students: Student-teacher anomaly detection with discriminative latent embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4183–4192.

Cohen, N.; and Hoshen, Y. 2020. Sub-image anomaly detection with deep pyramid correspondences. *arXiv preprint arXiv:2005.02357*.

Defard, T.; Setkov, A.; Loesch, A.; and Audigier, R. 2021. Padim: a patch distribution modeling framework for anomaly detection and localization. In *International Conference on Pattern Recognition*, 475–489. Springer.

Deng, H.; and Li, X. 2022. Anomaly Detection via Reverse Distillation from One-Class Embedding.

arXiv preprint arXiv:2201.10703.

Fort, S.; Ren, J.; and Lakshminarayanan, B. 2021. Exploring the limits of out-of-distribution detection. *Advances in Neural Information Processing Systems*, 34.

Kamoona, A. M.; Gostar, A. K.; Bab-Hadiashar, A.; and Hoseinnezhad, R. 2021. Anomaly Detection of Defect using Energy of Point Pattern Features within Random Finite Set Framework. *arXiv preprint arXiv:2108.12159*.

Kwon, G.; Prabhushankar, M.; Temel, D.; and AlRegib, G. 2020. Backpropagated gradient representations for anomaly detection. In *European Conference on Computer Vision*, 206–226. Springer.

Li, C.-L.; Sohn, K.; Yoon, J.; and Pfister, T. 2021. Cutpaste: Self-supervised learning for anomaly detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9664–9674.

Liang, Y.; Zhang, J.; Zhao, S.; Wu, R.; Liu, Y.; and Pan, S. 2022. Omni-frequency Channel-selection Representations for Unsupervised Anomaly Detection. *arXiv preprint arXiv:2203.00259*.

Liu, W.; Li, R.; Zheng, M.; Karanam, S.; Wu, Z.; Bhanu, B.; Radke, R. J.; and Camps, O. 2020. Towards visually explaining variational autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8642–8651.

Liu, Y.; Zhuang, C.; and Lu, F. 2021. Unsupervised Two-Stage Anomaly Detection. *arXiv preprint arXiv:2103.11671*.

Massoli, F. V.; Falchi, F.; Kantarci, A.; Akti, Ş.; Ekenel, H. K.; and Amato, G. 2021. MOCCA: Multilayer One-Class Classification for Anomaly Detection. *IEEE Transactions on Neural Networks and Learning Systems*.

Peng, X.; Wang, K.; Zhu, Z.; and You, Y. 2022. Crafting better contrastive views for siamese representation learning. *arXiv preprint arXiv:2202.03278*.

Perera, P.; Nallapati, R.; and Xiang, B. 2019. Ocgan: One-class novelty detection using gans with constrained latent representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2898–2906.

Pirnay, J.; and Chai, K. 2022. Inpainting transformer for anomaly detection. In *International Conference on Image Analysis and Processing*, 394–406. Springer.

Reiss, T.; and Hoshen, Y. 2021. Mean-shifted contrastive loss for anomaly detection. *arXiv preprint arXiv:2106.03844*.

Rippel, O.; Mertens, P.; and Merhof, D. 2021. Modeling the distribution of normal data in pre-trained deep features for anomaly detection. In *2020 25th International Conference on Pattern Recognition (ICPR)*, 6726–6733. IEEE.

Ristea, N.-C.; Madan, N.; Ionescu, R. T.; Nasrollahi, K.; Khan, F. S.; Moeslund, T. B.; and Shah, M. 2021. Self-Supervised Predictive Convolutional Attentive Block for Anomaly Detection. *arXiv preprint arXiv:2111.09099*.

Roth, K.; Pemula, L.; Zepeda, J.; Schölkopf, B.; Brox, T.; and Gehler, P. 2021. Towards total recall in industrial anomaly detection. *arXiv preprint arXiv:2106.08265*.

Rudolph, M.; Wandt, B.; and Rosenhahn, B. 2021. Same same but differnet: Semi-supervised defect detection with normalizing flows. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 1907–1916.

Rudolph, M.; Wehrbein, T.; Rosenhahn, B.; and Wandt, B. 2022. Fully Convolutional Cross-Scale-Flows for Image-based Defect Detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 1088–1097.

Schlüter, H. M.; Tan, J.; Hou, B.; and Kainz, B. 2021. Self-Supervised Out-of-Distribution Detection and Localization with Natural Synthetic Anomalies (NSA). *arXiv preprint arXiv:2109.15222*.

Sohn, K.; Li, C.-L.; Yoon, J.; Jin, M.; and Pfister, T. 2020. Learning and evaluating representations for deep one-class classification. *arXiv preprint arXiv:2011.02578*.

Song, J.; Kong, K.; Park, Y.-I.; Kim, S.-G.; and Kang, S.-J. 2021. AnoSeg: Anomaly Segmentation Network Using Self-Supervised Learning. *arXiv preprint arXiv:2110.03396*.

Tan, J.; Hou, B.; Batten, J.; Qiu, H.; and Kainz, B. 2020. Detecting outliers with foreign patch interpolation. *arXiv preprint arXiv:2011.04197*.

Wan, Q.; Gao, L.; Li, X.; and Wen, L. 2021. Industrial Image Anomaly Localization Based on Gaussian Clustering of Pre-trained Feature. *IEEE Transactions on Industrial Electronics*.

Yamada, S.; and Hotta, K. 2021. Reconstruction Student with Attention for Student-Teacher Pyramid Matching. *arXiv preprint arXiv:2111.15376*.

Yang, J.; Shi, Y.; and Qi, Z. 2020. Dfr: Deep feature reconstruction for unsupervised anomaly segmentation. *arXiv preprint arXiv:2012.07122*.

Yi, J.; and Yoon, S. 2020. Patch svdd: Patch-level svdd for anomaly detection and segmentation. In *Proceedings of the Asian Conference on Computer Vision*.

Yu, J.; Zheng, Y.; Wang, X.; Li, W.; Wu, Y.; Zhao, R.; and Wu, L. 2021. FastFlow: Unsupervised Anomaly Detection and Localization via 2D Normalizing Flows. *arXiv preprint arXiv:2111.07677*.

Zavrtanik, V.; Kristan, M.; and Skočaj, D. 2021. DRAEM-A discriminatively trained reconstruction embedding for surface anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8330–8339.

Zheng, Y.; Wang, X.; Deng, R.; Bao, T.; Zhao, R.; and Wu, L. 2021. Focus Your Distribution: Coarse-to-Fine Non-Contrastive Learning for Anomaly Detection and Localization. *arXiv preprint arXiv:2110.04538*.