International Journal of Machine Learning and Cybernetics (2024) 15:355–369
https://doi.org/10.1007/s13042-023-01913-7

ORIGINAL ARTICLE

RAMFAE: a novel unsupervised visual anomaly detection method based on autoencoder

Zhongju Sun[1] · Jian Wang[1] · Yakun Li[1]

Received: 5 January 2023 / Accepted: 25 June 2023 / Published online: 21 July 2023

Abstract

Traditional methods of visual anomaly detection based on reconstruction often use normal data to train autoencoder. Then the metric distance detection method is used to estimate whether the samples of detection belong to the exception class. However, this method has some problems that the autoencoder produces blurry images to cause false detection of normal pixel points. The model may still be able to fully reconstruct the undiscovered defects due to the large capacity of autoencoder, even if it is trained only on normal samples. Then, the metric distance detection method would ignore local key information. To solve this problem, this paper comes up with the random anomaly multi-scale feature focused autoencoder (RAMFAE), an innovative unsupervised visual anomaly detection technique, which incorporates three novel concepts. First, a multi-scale feature focused extraction (MFFE) network structure is designed and added between the encoder and decoder, which effectively solves the problem of reconstructing image blur and effectively improves the sensitivity of the model to normal regions. Second, this article employs Delete Paste, a novel data augmentation strategy for generating two different types of random anomalies, which pastes the cut part into a random location, while the pixels in the original position are filled with 0. In spite of the input anomalous images, the strategy makes the model be able to produce normal images to avoid the phenomenon of anomaly reconstruction, and then enables defect localization based on the error between the measured image and the reconstructed image. Third, the study adopts the image quality assessment with combining gradient magnitude similarity deviation (GMSD) and structural similarity (SSIM) to solve the problem that local key information and texture detail information are not easy to be paid attention to by the model, and alleviate the training pressure caused by Delete Paste enhancement. We perform an extensive evaluation on the challenging MVTec AD data set and compare it with the advanced visual anomaly detection methods in recent years as well. The AUC final result of RAMFAE in this text reaches 94.5, which is 3.6, 2.5 and 0.8 higher than the advanced IGD, FCDD and RIAD detection methods.

Keywords Anomaly detection · Data enhancement · Image quality assessment · Unsupervised learning · Multi-scale feature extraction · Attention mechanisms

1 Introduction

Vision-based industrial visual anomaly detection has become one of the emerging hot issues because of its important application value. In actual industrial production, normal products are relatively common, but abnormal products are extremely rare. Having labeled datasets is expensive, and samples from different production environments can vary depending on the shooting conditions, where unsupervised visual anomaly detection is more appropriate.

✉ Jian Wang
ganard@163.com

Zhongju Sun
toorud4c@163.com

Yakun Li
phiamlee@163.com

[1] College of Control Science and Engineering, Bohai University, Jinzhou 121000, China

In recent years, deep learning [4] has been well developed in the field of target detection and semantic segmentation, showing excellent advantages in image processing. And then researchers have applied deep learning to industry by proposing detection methods based on image reconstruction. Kang [19] proposed an anomaly detection algorithm for hyperspectral images based on initial detection based on attribute filtering and post-processing method based on edge-preserving filtering. Schlegl [27] is the first to introduce generative adversarial networks (GAN) [13] to defect detection, which makes randomly generated images mismatch with the samples to be tested by the adversarial idea. f-AnoGAN [28] introduces an additional encoder based on AnoGAN for feature extraction and takes a multi-stage training approach. However, GAN has a thorny problem in training, such as often training instability, gradient disappearance, and pattern collapse.

Based on the autoencoder (AE) [15], the network using encoder-decoder structure realizes the defect localization according to the reconstruction error between the input image and the reconstructed image. However, AE will have blurry phenomenon in reconstructing images. Bergmann [1] examines datasets of woven fabrics and nanofibers with applying the SSIM [35] to AE. Although the authors' own network structure produced clearer results, it is not robust enough for feature extraction and more difficult to learn for complex samples. Zavrtanik [42] proposed his evaluation method multi-scale gradient magnitude similarity (MSGMS) and completed the visual anomaly detection and localization with using image restoration to avoid the reconstruction of errors. However, this method divides each image into three masks, which requires more training time and produces false detection during positioning as well.

In order to ensure the clarity of the reconstructed image and the robustness to complex samples, a new visual anomaly detection method RAMFAE is proposed in this paper. The contributions of this work are as follows:

1. Compared with GAN, training procedure of AE is more stable, but it takes more training time to generate clear images. The ability to reconstruct images is less potent, and even the generated images may be blurry. Therefore, this paper designs a deeper network structure RAMFAE, in which the MEEF module can be used to extract multi-scale features, and the attention mechanism makes the model pay more attention to the normal region and thus be more sensitive to the normal region. At the same time, the anomalous area is also poorly reconstructed to reduce the abnormal reconstruction. In the study residual module is used in the decoder module, which improves the performance of the model.
2. Due to its strong generalization ability, even if the traditional AE uses normal pictures as training samples, it often reconstructs unseen anomalies. In this paper, we employ a novel enhancement—Delete Paste that can produce two different types of anomalies in typical samples in training. This technique involves pasting the cut portion to a random location while filling the pixels at the original location with 0. The strategy improves the robustness of the model, generating normal pictures even in the abnormal data, and then achieving defect localization according to the error between the image to be measured and the reconstructed image.
3. For simplicity and speed, the method of measuring the distance between pixels is usually selected to calculate the error. However, the autoencoder needs to obtain the average value when calculating the loss of the original image and the reconstructed image, and the local key information will be erased. Especially in areas with rich image texture. In this paper, we use the image quality assessment (IQA) that combines GMSD and SSIM. This makes the proposed model pay more attention to the local gradient field and texture features of the image, reconstructing the image details accurately, and relieving the training pressure caused by superimposing the mask on the original image. GMSD avoids the information averaging caused by averaging by calculating the feature information of each local gradient field. SSIM pays more attention to the significant visual changes made by the autoencoder, so as to achieve accurate segmentation of defects. During the test, additional MSGMS was used to calculate the loss at the pixel level, which was confirmed by the researchers to have excellent results [42].

The rest of this paper is structured as follow. In Sect. 2, this paper discusses the work related to visual anomaly detection, while in Sect. 3, the visual anomaly detection method of RAMFAE is described in detail. Next, Sect. 4 describes the design of the experiments in this paper, and showing the results of RAMFAE in visual anomaly detection, ablation experiments, and IQA applications. Finally, the article is summarized in Sect. 5.

2 Related work

For model training, unsupervised deep learning-based techniques just require normal samples. It is not only fixing the issue that supervised deep learning cannot detect unknown defects, but also has stronger feature representation than traditional one. The core of this method is to achieve the detection and localization of defects based on the difference in pixels or features. The image-based reconstruction method just trains the model on normal samples. Since the parameters of the model are just trained by normal samples, the model can only reconstruct normal samples well, and the reconstruction error will be large in the defect region of defective samples.

These techniques may employ the autoencoder models, like AE. However, AE has a blurry phenomenon during reconstruction, so it is easy to cause false detection of normal pixels when calculating the reconstruction error. Yang [40] enhances the accuracy and clarity of the reconstructed images by extracting multi-scale feature information and offering various fine-grained contextual information. However, it requires 700 epochs of training time to reconstruct high quality images and employs VGG19 [30] pre-trained on ImageNet [26] to generate hierarchical image features. Although the authors focused on multi-scale feature information, they neglected the effect of deep network structure on feature extraction ability. Liu [24] performs multi-scale fusion on the features of regions and boundaries, which provides an effective solution for boundary recognition of low-contrast medical images. Zahra [41] uses supplementary distribution in the input space to separate the abnormal samples from the reconstruction, which reduces the reconstruction error. Conversely, Chung [7] makes good use of the blurring effect by stylizing the input image, transforming the original image into an image with the same style as the reconstructed image. To extract features for images and produce high-quality images, we employ a deeper network structure and the MFFE in this paper, which highly can extract features from the feature map of potential space. And the attention mechanism is used for feature maps extracted by multi-scale features to make the model pay more attention to the anomaly-free areas. The Residual module is added to the decoder module in the article, which is able to cross links, weaken the strong connection between each layer to avoid the problem of deep network gradient disappearance and slow convergence.

The hypothesis of the image reconstruction method is not completely reliable. Training only on normal samples can reconstruct normal images well, but the model may still reconstruct unknown defects completely. The researchers found that after adding defects to normal images by using data augmentation, training the network model can restore it to the corresponding original image. After training is complete, the model has the ability to eliminate defects based on context. The test phase uses the reconstruction error of the recovered image and the input image to perform defect segmentation.

Data augmentation improves the training effect and generalization of the model from the perspective of training data. By making full use of the given data and the characteristics of the defect to design corresponding defect data to be superimposed with the normal sample to expand the defect sample. Tao [33] generated defect samples for the copper wire data of deep-hole parts using affine transformation, Gaussian blurring, and pretzel noise. Yang [39] enhanced the defect samples for fabric data by producing random masks overlaid with normal samples to mimic actual defect patterns. Lin [23] augments defective data for multi-category industrial product data by removing defective parts from defective samples and fusing them with healthy samples. By copying and pasting other normal regions, Li [21] creates negative samples that can be used to train supervised classification agents. On the basis of CutPaste [21], Schluter [29] used Poisson image editing to seamlessly mix multiple image blocks as synthetic defects, which improved the diversity and authenticity of defects. In this study, we develop the Delete Paste crop method, which can simultaneously produce two different exception forms. The original pixels of the mask region are cropped and stitched to random positions, while the black mask region is generated randomly. The network model can recover anomalous images to the corresponding original images by this strategy, which also improves the ability of model to generalize. Then defect detection is accomplished according to the error between the measured image and the reconstructed image.

Although the data enhancement approach can enhance the generalization of model, it also makes model training more challenging. The input image enhanced by Delete Paste has a mask of a larger area and pixel information spliced from an unknown area. It is difficult to estimate the gap between the original image and the reconstructed image only from measuring the distance. At this time, human perception should be used to estimate the pixel gap between images, rather than relying on a single indicator. Image quality assessment (IQA) primarily uses characterization studies of images to assess the image merit and the distortion level. The input image is transformed into a grayscale image by CG-DIQA [22], scaled to a fixed size, and then blocks of candidate characters are found using the MSER method. Next, the standard deviation of the gradient of the candidate characters is calculated, and finally, the quality score of the document image is estimated. The gradient similarity-based method MSGMS proposed by Zavrtanik [42] was used for training and anomaly score estimation with better results. In this paper, the IQA method based on GMSD [37] and SSIM [35] is used as the loss function in the training phase. They are designed based on human senses, making the model more sensitive to local key information and graphic texture information. Accurate reconstruction of the details of the image is helpful for training problems caused by Delete Paste. The MSGMS was additionally used as an anomaly evaluation strategy during testing, and it experimentally proved to be useful.

3 Methods

The proposed RAMFAE consists of three parts: Delete Paste, MFFE, and IQA. The main process of the method is shown in Fig. 1. The first step of RAMFAE is Delete Paste. In order to avoid the autoencoder may be generalized to defects, and even the input data and output data tend to be consistent, this paper proposes Delete Paste. To simulate real defects, Delete Paste uses pixels in the random area of the sample as a mask. Paste the mask to other areas of the image and delete the pixels in the original area. After adding defects to the normal image, the training network model restores it to the corresponding original image, which can avoid the problem of identity mapping to a certain extent.

The normal image with defects is added as the input data into multi-scale feature focused autoencoder. This step is to obtain the characteristics of multi-scale semantic information. The feature layer information from encoder is extracted on different receptive fields, and the obtained feature layer ME has rich semantics. The second part of the MFFE is coordinate attention, which aggregates the vertical and horizontal input features into two independent direction-aware feature maps. The attention mechanism makes the network more sensitive to normal areas, but not to abnormal areas. To avoid the problem of deep network gradient disappearance and slow network convergence, the residual module is added to the decoder part.

In the training stage, this paper hopes that the reconstructed image is normal. Under the action of IQA, the local area, and texture feature part are accurately reconstructed, and the loss between the original image and the original image is minimized. In the test, the Gaussian smoothing with a step size of 21 is used to extract the abnormal score. The Gaussian smoothing operation is a two-dimensional convolution operation for removing noise from blurred images. And the abnormal distribution is displayed on the original image in the form of a heat map. In the following, we describe our method in detail.

3.1 Multi-scale feature focused autoencoder

The RAMFAE network structure is depicted in Fig. 2. Encoder, MFFE, and decoder are the three components of the method. The RAMFAE network serves as the backbone and the general encoder-decoder structure. In comparison to SSIM, RAMFAE employs a deeper convolutional neural network, which has a stronger capability for feature extraction and speeds up convergence. The RAMFAE in this study includes, similar to Table 1, four layers of EncoderBlock, three layers of DecoderBlock, a layer of Conv to shrink the channel to the dimensions of picture. Finally, Tanh is used as the activation function. MFFE is

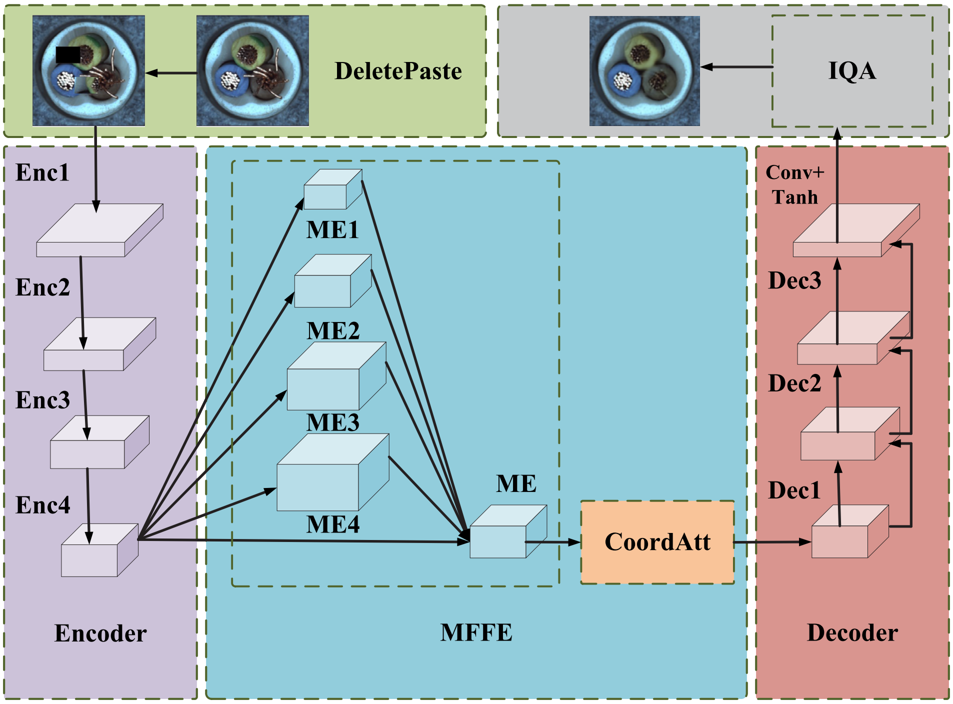

Fig. 1 The input normal image is processed by Delete Paste to become data with anomalies. The loss between the image reconstructed by MFFAE and the original image is calculated by image quality assessment

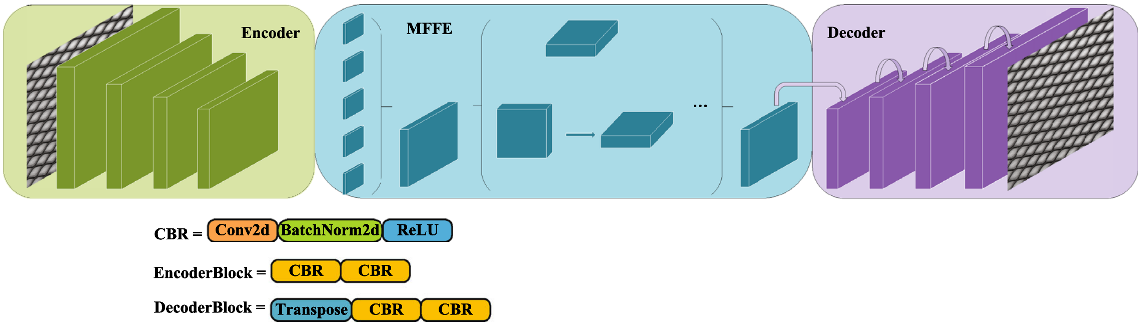
Encoder
MFFE
...
Decoder
CBR = Conv2d BatchNorm2d ReLU
EncoderBlock = CBR CBR
DecoderBlock = Transpose CBR CBR

Fig. 2 The original image is reconstructed by RAMFAE, and EncoderBolck consists of two CBRs, to which DecoderBlock adds transposed convolution for the upsampling. The CBR module is composed of Conv, batch normalization, and Relu activation functions. MFFE is used in between the encoder and decoder to extract multi-scale information and focus on the normal region.The residual structure is used between layers of decoder to let some layers of the neural network skip the connection of neurons in the next layer and weaken the strong connection between each layer

Table 1 Structure of the RAMFAE network model

Stage	Operator	Stride	Channels	Padding
0	EncoderBlock3×3	1	64	1
1	EncoderBlock4×4	2	128	1
2	EncoderBlock4×4	2	256	1
3	EncoderBlock4×4	2	512	1
4	MFFE	#	512	#
5	DecoderBlock3×3	1	256	1
6	DecoderBlock3×3	1	128	1
7	DecoderBlock3×3	1	64	1
8	Conv3×3	1	3	1
9	Tanh	#	#	#

Three layers of DecoderBlock, four layers of EncoderBlock, a layer of Conv to shrink the channel to the size of the image, and Tanh as the activation function round out the structure. MFFE is used to perform feature extraction and feature focus between the encoder and decoder

used to perform feature extraction and feature focus between the encoder and decoder.

The purpose of the encoder component is to convert the feature data of original image into a feature map of the potential space. Two CBRs make up each EncoderBolck layer to improve feature extraction. The usual implementation of each convolutional layer includes convolution, batch normalization (BN) [18], and rectified linear unit (ReLU) [10].

For both normal and abnormal regions, the unprocessed latent has a lot of randomness. In order to enhance the quality of the reconstructed images and make the model more sensitive to normal regions, which avoids the reconstruction of abnormal regions, the multi-scale feature-focused extraction (MFFE) is used between the encoder and decoder to perform deep multi-scale feature extraction for the feature map of potential space.

The decoder takes the potential low rank space-processed vectors and extracts the information from them, processing them as vectors of the original image size. The extraction of features using two layers of CBR is necessary for the decoder block, but also an upsampling module is needed. The scaled high-frequency component of image is lost due to the low-pass filter nature of the bilinear interpolation method, and the edges of the image become somewhat more blurry. Transposed convolution provides the best results with the least amount of post-processing image quality loss, smoother edges than bilinear interpolation, and high computational accuracy. As the upsampling module in this study, we employ a transposed convolution with a kernel size of 4, stride size of 2, and padding of 1.

This paper discovers that due to the increase of network depth during the training process causes degradation of the weight matrix, which made the model hard to train. In this study, we employ residual networks [34] in the decoder module, which permits some neural network layers to connect to every other layer without connecting the neurons in the layer below, weakening the strong connection between each layer. Residual effectively solves the problems of difficult convergence and slow learning of the model that appears in the experiments and improves the performance of the model. And residual is a plug-and-play module that does not introduce redundant parametric quantities and requires only simple modifications to improve the performance of AE, which is worth promoting in the field of visual anomaly detection.

3.1.1 Multi-scale feature focused extraction module

To reconcile the conflict between the competing goals of having a larger receptive field for the features extracted from the image and maintaining a reasonable resolution for the feature map, dilated convolution is employed.

Intensifying the receptive field: the receptive field is expanded with conventional downsampling, but the spatial resolution is decreased. Dilated convolution can be used to increase the receptive field while ensuring resolution. Capture context data on multiple scales: The expansion rate, which regulates the padding and dilation during convolution, distinguishes the dilated convolution layer from the general convolution. Different scale receptive fields can be obtained through different filling and expansion, and multi-scale information can be extracted.

The multi-scale extraction part of MFFE extracts feature maps of different receptive fields from the feature map of the Encoder. As shown in Fig. 3, the feature extraction module used in this paper consists of three parts. The top is a 1 × 1 convolution to reduce the dimension of the original data. But the output dimension of this paper is set to the same as the input dimension, retaining the original semantic information. The middle three layers are pooled pyramids, using dilated convolutions to enlarge the receptive field. If the expansion rate of dilated convolution is set too large, a large number of parameters will be added, which will increase the difficulty of training. In this experiment, the expansion factor is set to 5, 7, 9. The lowermost Pooling layer is first an AdaptiveAvgPool layer. The features of each channel are extracted, and the global features are obtained by compressing the feature maps of each channel to 11. The features obtained in the previous step are then further extracted and their dimensions are decreased using a 1 × 1 convolution layer. Create a pyramid for pooling whereupon. The corresponding dilated convolution layers are superimposed to extract features at various scales for a given expansion factor atrous rates. After a 1 × 1 convolution, the feature layers at various scales are combined to produce the final result.

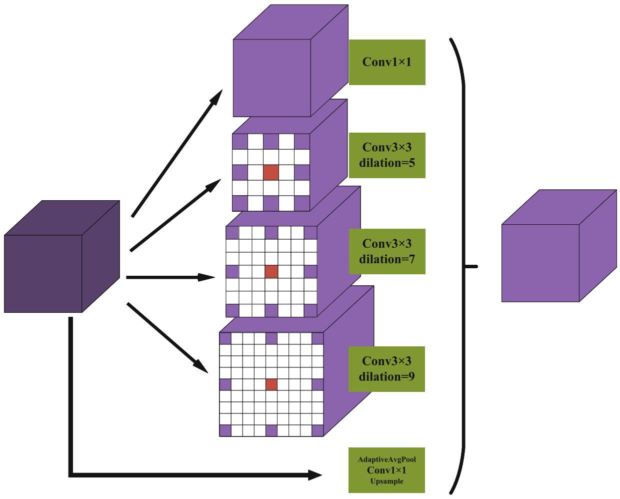

Fig. 3 For a given expansion factor, the corresponding dilated convolution layers are superimposed to extract features at different scales. Add an empty pooling layer and convolve the output of the stacked layers to get the final result

The semantic information in the feature layer is currently rich. In order to reduce the randomness of the latent space to the normal and abnormal regions, the attention mechanism is introduced to make the model pay more attention to the normal regions in the anomaly samples.

The most popular attention mechanism is still SE attention proposed by SENet [17], which calculates channel attention through 2D global pooling, providing significant performance improvement at a fairly low computational cost. However, the SE only considers the encoding of inter-channel information and ignores the importance of position information, which is actually essential for many visual tasks with capturing the target structure. Later, CBAM [36] etc. used large-size convolution to utilize location information by reducing the number of channels. Nevertheless, CBAM convolution can only capture local correlations. For spatial dependencies that are important for visual tasks, modeling does not achieve the desired results.

This paper follows the multi-scale extraction part with coordinate attention [16] to address the aforementioned issue. Figure 4 illustrates how the vertical and horizontal input features are combined into two independent direction-aware feature maps using two one-dimensional global pooling operations. Then, two attention maps are encoded from the two feature maps with specific direction information, each of which captures the feature information of the input feature map along a spatial direction. Specifically, for the input X, each channel is first encoded in the horizontal and vertical coordinate directions using pooling cores of sizes (H, 1) and (1, W), so that the output of the c channel at height h is expressed as follows:

$$z_c^h(h) = \frac{1}{W} \sum_{0 \leqslant i < W} x_c(h, i). \tag{1}$$

The output of the c channel with width w is expressed as follows:

$$z_c^w(w) = \frac{1}{H} \sum_{0 \leqslant j < H} x_c(j, w). \tag{2}$$

These two transformations also allow the attention module to capture precise location information along two different spatial directions, which helps the network more accurately locate the target of interest. Then, the two feature maps are cascaded, and then $F1$ transform is performed using a shared 1 × 1 convolution. The generated f is an intermediate feature map for spatial information in the horizontal and vertical directions.

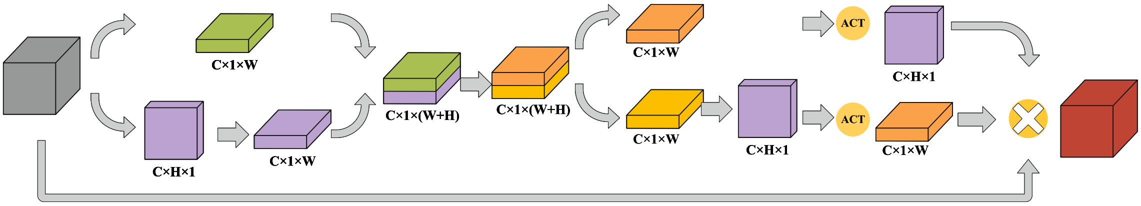

Fig. 4 Network structure diagram of coordinate attention mechanism

$$f = \delta(F_1([z^h, z^w])). \tag{3}$$

The f is divided into two separate tensors f^h and f^w along the spatial dimension, and then the feature map is transformed to the same number of channels as the input X by using two 1×1 convolutions F_h and F_w.

$$\begin{aligned} g^h &= \sigma(F_h(f^h)) \\ g^w &= \sigma(F_w(f^w)) \end{aligned} \tag{4}$$

The final output of the coordinate attention module can be expressed as follows. Therefore, the coordinate attention module completes both horizontal and vertical attention, and it is also a channel attention.

$$y_c(i,j) = x_c(i,j) \times g_c^h(i) \times g_c^w(j). \tag{5}$$

3.2 Delete paste

In the process of semantic segmentation and object detection, data augmentation is a crucial step. The most popular techniques are Cutout [9], MixUp [43], Mosaic [3], GridMask [5], CutPaste [21], illumination distortion, geometric distortion, random erasure, etc. The data augmentation method can be used in visual anomaly detection to create similar abnormal parts in the original image. This study aims to produce normal images from abnormal images.

Although popular methods like rotary cutting [11, 14, 31] have been studied in the context of visual anomaly detection, simply using existing methods for defect detection is not optimal. For all geometric transformations of translation and rotation are helpful for learning sample semantics, the transformations of this method are too regular. This paper designs a data augmentation strategy, which can provide variable abnormal effects. These abnormal effects are encountered during training, and they are removed and restored to a normal appearance, and it is expected that the existing abnormalities will be cleared in the test.

The cutout is an effective data expansion method. He cuts a small block in the region to generate a rectangular mask with improving the accuracy of the image classification task, but the effect of this method is single. This paper designs the Delete Paste enhancement from Cutout, it encourages the model to learn variable abnormal effects to avoid defect reconstruction.

Initialize an M with all pixel values of 1. M_{S_i} is a mask with 0 pixels in S_i region. By multiplying M_{S_i}, the region belonging to S_i in I is set to zero, and the pixel value not belonging to S_i region is unchanged. The pixels before being filled are pasted into other areas of the image. As shown in Fig. 5, through a Delete Paste enhancement, two different types of anomalies can be obtained, making the sample more irregular.

3.3 Image quality assessment

Image quality assessment, also known as IQA, plays an important role in digital image science. It involves not only physical imaging from the natural world to digital images, but also human psychological and physiological perception of signals.

Gradient intensity similarity deviation, also known as gradient magnitude similarity deviation (GMSD) [37], can reflect the structural information of the image. The local image quality is evaluated by calculating the similarity of local gradient intensity, and the standard deviation of local

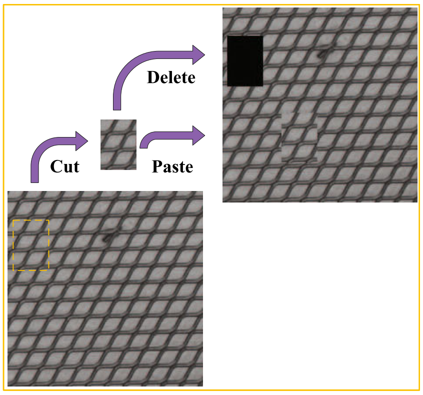

Fig. 5 Delete paste enhancement process. First, a rectangular region is randomly cut from the original sample. The mask M_{S_i} is set in the coordinates of the rectangular region. By multiplying M_{S_i}, the region belonging to S_i in I is set to zero, and the pixel value not belonging to S_i region remains unchanged. The original pixel value of this rectangle randomly moves to other positions of the sample

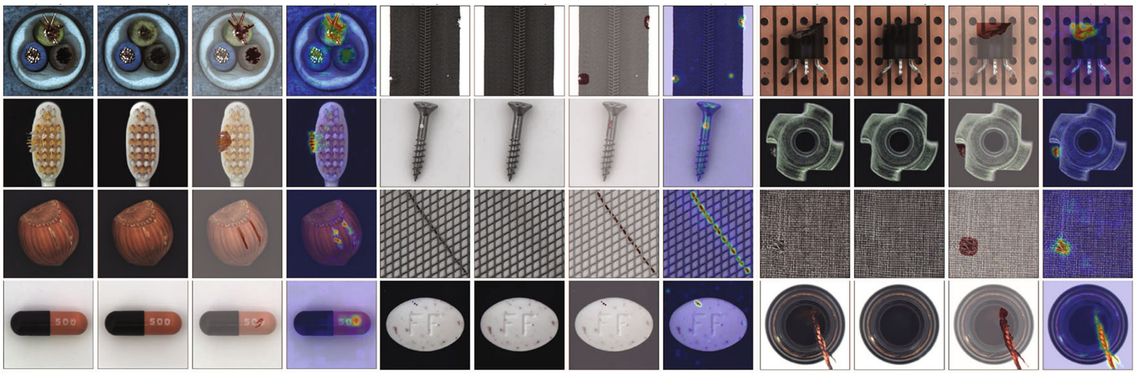

Fig. 6 Cable, zipper, transistor, toothbrush, screw, metal nut, hazelnut, grid, carpet, capsule, pill, bottle in MVTec AD data set. Each sample has four images representing the original image, the reconstructed image, the ground truth position in brown, and the Heat Map using a pixel-level detector

image quality is calculated to measure the quality of the whole image.

Calculate the similarity of local gradient strength, using the Sobel operator. The Roberts operator to obtain gradient information, and choose to use the standard 3 × 3 Prewitt operator.

$$h_x = \begin{bmatrix} \frac{1}{3} & 0 & -\frac{1}{3} \\ \text{[7}pt\text{]}\frac{1}{3} & 0 & -\frac{1}{3} \\ \text{[7}pt\text{]}\frac{1}{3} & 0 & -\frac{1}{3} \end{bmatrix}, h_y = \begin{bmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \text{[7}pt\text{]}0 & 0 & 0 \\ -\frac{1}{3} & -\frac{1}{3} & -\frac{1}{3} \end{bmatrix} \tag{6}$$

To obtain the gradient intensity of the image in the horizontal and vertical directions, denoted as $m_r(i)$ and $m_d(i)$ where i is the position of the pixel point, the operator is convolved with the image I

$$\begin{aligned} m_r(i) &= \sqrt{(r \times h_x)^2 + (r \times h_y)^2} \\ m_d(i) &= \sqrt{(d \times h_x)^2 + (d \times h_y)^2}, \end{aligned} \tag{7}$$

where $m_r(i)$ and $m_d(i)$ are the gradient intensity maps of the image, $*$ is the convolution symbol. The gradient magnitude similarity (GMS) is defined as:

$$GMS(i) = \frac{2m_r(i)m_d(i) + c}{m_r^2(i) + m_d^2(i) + c}. \tag{8}$$

A constant named c is set to prevent the denominator from being 0. Here, it is possible to average each local gradient field; this process is known as GMSM:

$$GMSM = \frac{1}{N}\sum_{i=1}^{N} GMS(i). \tag{9}$$

The following formula, combines and uses the change in the quality of local regions in the global image response as a loss function to preserve the original semantics of pixel points, while also taking into account the fact that natural images typically have multiple local structures and that different local regions averaged directly will be affected more and lose the original rich semantics:

$$GMSD = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(GMS(i) - GMSM)^2}. \tag{10}$$

This approach is similar to the averaging pooling method of SSIM. In the testing phase, this paper uses pixel metrics to score the anomalies and takes the averaging operation in the channel dimension and uses the following formula:

$$L_{GMSD} = \frac{1}{N}\sum_{i=1}^{N}(GMS(i) - GMSM)^2. \tag{11}$$

The GMS is extended to a multi-scale variant MSGMS by computing it on multiple image scales. The MSGMS loss is defined as the average of the GMS distance map at multiple scales:

$$L_{MSGMS} = \frac{1}{4}\sum_{l=1}^{4}\frac{1}{N}\sum_{i=1}^{H}\sum_{j=1}^{W} 1 - GMS(x, y)_{i,j}. \tag{12}$$

MSGMS is also used in the test phase where l is the loss of image pyramid at four different scales. SSIM is another metric of similarity between images. Given two images x and y, the structural similarity between the two images is calculated according to the following formula:

$$\begin{aligned} SSIM(x, y) &= \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \\ L_{SSIM} &= \frac{1}{N}\sum_{i=1}^{H}\sum_{j=1}^{W} 1 - SSIM(x, y)_{i,j}. \end{aligned} \tag{13}$$

Where μ_x and μ_y are the mean values of image x and y, respectively, σ_x^2 and σ_y^2 are the variance of x and y, and σ_{xy}

is the covariance of x and y. $c_1 = (k_1 L)^2$, $c_2 = (k_2 L)^2$ are the constants used to maintain stability. L is the dynamic range of pixel values, taken as 0.01 and 0.03, respectively. Since SSIM is used as the loss function, no changes in luminance and contrast are selected for data enhancement. Finally, combined with the MSE loss, the total loss function is defined as:

$$L = \lambda_1 L_{GMSD} + \lambda_2 L_{SSIM} + \lambda_3 L_2. \tag{14}$$

The uniqueness of the proposed method lies in a novel RAMFAE algorithm, utilizing Delete Paste to obtain two types of irregular anomalies, and exploiting multi-scale feature extraction, coordinate attention mechanism and hybrid image quality assessment to play a role in network feature information flow. Compared to using raw images without processing, the Delete Paste strategy makes the generalization ability of model improved and the input anomalous data can generate normal images better. MFFE solves the problem of blurred generated images in AE and does not produce false detection of normal pixel points. The difficulty of training the data processed by Delete Paste as reflected in the experiments can be solved by GMSD and SSIM which are sensitive to local gradients and texture details. The above methods' ablation tests will be covered in Sect. 5.

4 Experiment

4.1 Experimental settings

In this paper, RAMFAE is used to test and evaluate the challenging MVTec AD [2] data set. Compared with other advanced localization methods AnoGAN [27], AE-SSIM [1], RIAD [42], DAGAN [32], SCADN [38], P-NET [44], IGD [6], FCDD [25], better results are obtained.

Data set: The MVTec AD data set contains 5354 high-resolution color images belonging to 15 categories of object and texture types. It has a total of 73 types of defects, such as scratches, dents, contamination, etc. It contains normal data set for training and abnormal data set for testing. Ground truth labels are also provided for pixel-level evaluation. The texture types contain carpet, grid, leather, tile, wood. Bottle, cable, capsule, hazelnut, metal nut, pill, screw, toothbrush, transistor, zipper belong to object types. For MNIST datasets, this includes 60,000 handwritten digits from ’0’ to ’9’. Each of these 10 numbers is used as the target class. The training data consists of 80% normal data, and the test data consists of the remaining 20% normal data and 10–50% abnormal data randomly selected from other categories. Therefore, the model only uses normal data for training and uses normal and abnormal data for testing.

Implementation details: This paper conducts experiments on 16-core Intel (R) Xeon (R) Platinum 8350C, RTX 3080 Ti 12 G memory. Delete Paste is used to generate exceptions during training. The optimizer selects Adam and sets the learning rate to 1e−4. Adjust the learning rate using CosineAnnealingLR, with the minimum learning rate fixed at 1e−6. Each experiment was trained for 250 epochs at image size 256 and the batch size was set to 16. AE-SSIM, RIAD, DAGAN, SCADN, P-NET, IGD also use Adam as the optimizer, and the learning rate is set to 1e−4.Each experimental model trains 250 epochs under the condition of image size of 256, where DAGAN trains 20,000 epochs and P-NET trains 800 epochs. The potential spatial dimensions of AE-SSIM and IGD are 500 and 128, respectively.

Metrics: The pixel-level evaluation metrics can be considered as the representation of the picture-level classification metrics in the pixel level. A ROC AUC score is calculated for each image by pixel, and an average score is calculated for the ROC AUC size of all images. In this case, the problem is a pixel-by-pixel binary classification problem, and thus can be evaluated using a pixel-level confusion matrix.

Baselines: AE-SSIM is based on CAE method. An anomaly is identified by pixel-by-pixel comparison between the original image and the reconstructed image. Training loss using SSIM.

AnoGAN is a GAN-based model. Its first attempt is to generate the nearest normal image of the test image using a GAN generator trained on normal data only. Then anomalies are detected by computing the per-pixel residuals between the test image and its nearest normal counterpart.

RIAD uses U-Net as the benchmark model. The input image is divided into many grids of $k \times k$ to create n masks. The grid pixel value of the mask is set to 0, and sent to the network for restoration. Multi-scale gradient magnitude similarity is used to calculate the loss of reconstruction results and input images.

The whole structure of DAGAN consists of two autoencoders. Inspired by U-Net, the generator is designed as an automatic encoder with jump connection structure. The jump structure provides good reconstruction ability. The discriminator is used to receive the reconstructed image and identify the difference between the original image and the reconstructed image.

Based on the autoencoder, SCADN adds a discriminator loss to help network reconstruction better. The middle layers use dilated convolution. For the input image, strip masks with different widths and different directions are randomly used for coverage, and then the image is restored.

The encoder of IGD is used to convert the image into representation z, and the decoder is used to reconstruct the image. MS-SSIM and MAE were used as training losses. Gaussian anomaly classifier and evaluator are added between encoder and decoder. The Gaussian anomaly classifier distinguishes training samples based on the distance from the training sample to the Gaussian center and the standard deviation of these distances.

P-Net consists of a structure extraction module, image reconstruction module, and structure extraction module. The structure information is extracted from the original image, and then the original image and the structure information are encoded respectively. The coding features are spliced and sent to the decoder for image reconstruction. The reconstructed image will pass through the structure extraction module again, and the reconstruction error and structural information error will be constrained.

FCDD uses FCN to map the image to the feature matrix. Only the convolution layer and pooling layer are used. A feature map is obtained by fully convolutional data description, and then the original size anomaly area interpretation map is obtained by upsampling. This paper proposes a heatmap upsampling algorithm. The center of the convolution kernel is a mean point of Gaussian distribution, and other points decrease with distance.

As a variant of AE, the main change of VAE variational autoencoder is the generation of coding. Further variational processing is performed on the autoencoder model so that the output of the encoder can correspond to the mean and variance of the target distribution.

MemAE adds a memory module to autoencoder. Given the input, the coding result is used as a query to retrieve the most similar item in the Mem module for reconstruction. In the training phase, the content in the Mem module is updated to construct the prototype element of the normal sample; in the test phase, the content of the Mem module is fixed, and the reconstruction is performed according to

Table 2 ROC AUC scores on hazelnut using different strategies

MFFE	Delete Paste	SSIM	GMSD	Score
				0.929
✓				0.958
	✓			0.963
✓	✓			0.967
		✓	✓	0.955
	✓	✓	✓	0.972
✓		✓	✓	0.958
✓	✓	✓	✓	0.985

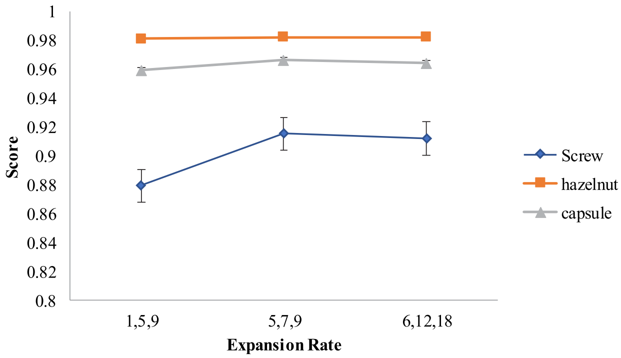

Fig. 7 ROC AUC scores for different expansion rates on data set

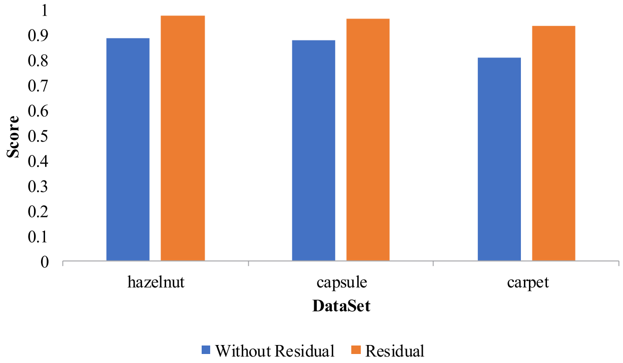

Fig. 8 Performance of residual structure in different datasets

Table 3 ROC AUC metrics on different image quality assessment strategies MVTec AD

Class	GMSD [37]	SSIM [35]	MSGMS [42]
Bottle	0.923	**0.952**	0.943
Cable	0.893	0.936	**0.940**
Capsule	**0.964**	0.940	0.961
Carpet	0.907	**0.937**	0.928
Grid	0.821	**0.942**	0.936
Hazelnut	0.953	0.961	**0.976**
Leather	0.948	**0.960**	0.956
Metal nut	0.906	0.892	**0.911**
Pill	0.974	**0.980**	0.978
Screw	0.879	0.912	**0.915**
Tile	0.903	**0.963**	0.950
Toothbrush	0.944	0.956	**0.959**
Transistor	0.917	0.921	**0.931**
Wood	0.865	0.872	**0.902**
Zipper	0.925	0.929	**0.947**
Avg	0.914	0.936	**0.942**

The bold values represent the best results of the three image quality assessments

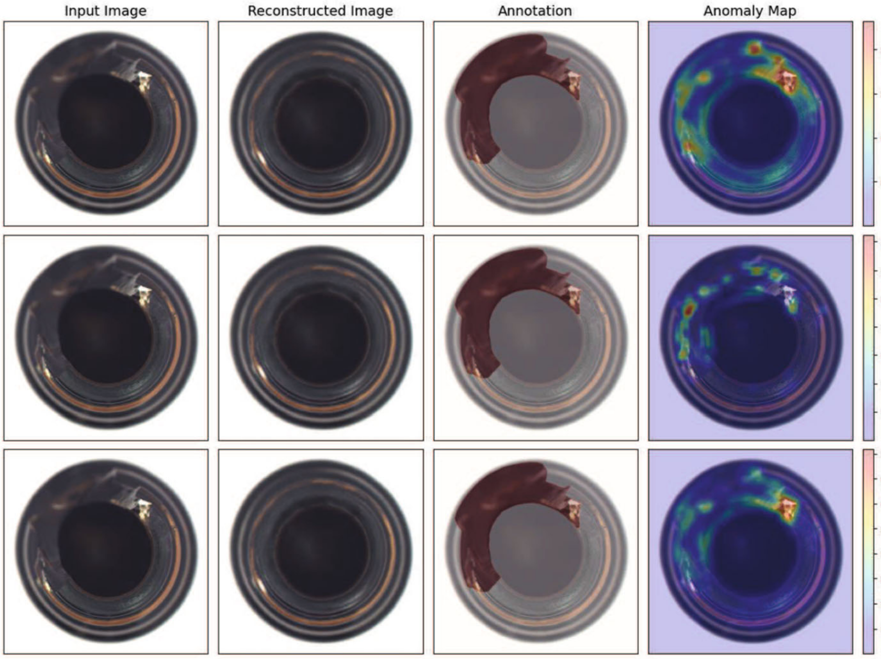

Fig. 9 The bottle in the MVTec AD data set is the heat map of anomaly detection generated by SSIM, GMSD, and MSGMS from top to bottom. The effect of SSIM anomaly detection is the best, but there is a small amount of false detection. The MSGMS detection area is concentrated in the anomaly area, and the GMSD anomaly hit rate is low

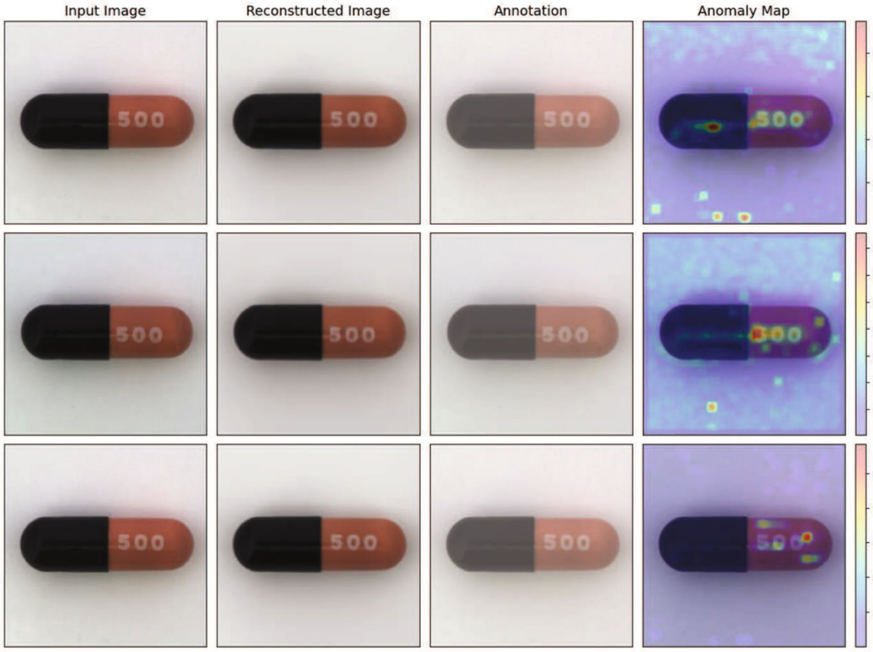

Fig. 10 From top to bottom are the heat maps of anomaly detection generated by SSIM, MSGMS and GMSD respectively. The area of false detection is larger in SSIM and MSGMS, and smaller in GMSD

Mem, and the reconstruction error is calculated for anomaly detection.

4.2 Ablation experiment

In this paper, ablation experiments are performed on the hazelnut data set. Table 2 shows the evaluation of the effects of the MFFE, Delete Paste model, and SSIM, GMSD image quality assessment strategies on the model using the AE as the backbone network. The MFFE, Delete Paste, SSIM and GMSD were added to the original AE. Without the use of modules, the anomaly detection on the hazelnut data set is 92.9. This article uses this as a benchmark score. The MFFE and Delete Paste modules and image quality assessment strategy were added to the experiment, resulting in an improved ROC AUC of the model at the pixel level. Afterwards, optimization modules

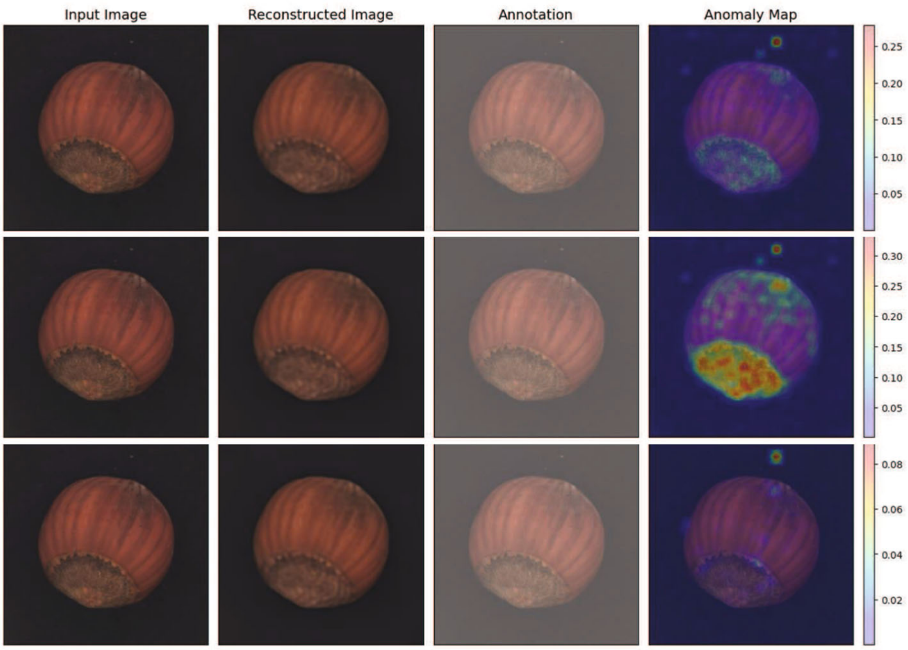

Fig. 11 From top to bottom, the heat maps of anomaly detection are generated by MSGMS, SSIM and GMSD respectively. GMSD and MSGMS show less abnormal areas, thus reducing the probability of false detection

were applied to other data sets and both showed significant results.

For dilated convolutions, a larger expansion rate can have a larger receptive field, but the training parameters will also increase. In this paper, three groups of different expansion rates are used for experiments, which are (1, 5, 9), (5, 7, 9), and (6, 12, 18). It is easy to know from Fig. 7 that the second group of experiments obtained the highest score. The first group had the lowest experimental score and the second and third groups had similar scores. However, it can be found in the experiment that the maximum expansion rate determines the training time. The third group of experiments greatly increased the training time due to the expansion rate of 18, even if a higher score was obtained. The first group and the second group had almost the same and fast training time. But the second group provided a larger receptive field, so the optimal ROC AUC results were achieved. Finally, (5, 7, 9) is used as the coefficient of dilated convolution.

This paper conducts ablation experiments on the Residual structure of the Decoder module on the hazelnut, capsule and carpet datasets. Hazelnut is a large random data set, which is difficult to train, slow to train and hinders convergence. Capsule is a moderate data set, easy to train, but the abnormal colors are not easy to detect as they are similar to normal colors. Carpet is a difficult textured data set and has failed to achieve the desired results in many approaches. The experimental results are shown in Fig. 8, where the method using the Residual structure works better than the method without the residual structure. Residual successfully addresses the issue of the model, like slow learning and difficult convergent behavior in the experiment, enhancing the performance of model.

4.3 Image quality assessment strategy selection

Since the model was trained to minimize both GMSD and SSIM, both were chosen as the anomaly score estimation function. In addition, MSGMS as a modified method of GMSD was proven to possess excellent results, so it was also used as the estimation function in this paper. The results of using GMSD, SSIM and MSGMS as the estimation function for the anomaly scores are shown in Table 3. Overall MSGMS is able to demonstrate better robustness on different types of data and achieves the highest overall score. However, SSIM shows a greater advantage on the textured data set.

The evaluation strategy used in SSIM can more effectively locate texture-like or dense anomalies. By emphasizing edge and texture similarity to identify anomalies, such as the fabric on the side of a zipper, wood, tile, carpet, and grid, it can mimic human perception. Then, for large defects SSIM is able to locate the largest area of anomalies, involving the widest range of colors, but the corresponding noise will cover more. MSGMS locates relatively less area, but the more important parts are highlighted. GMSD locates the least area, and only the parts with the largest quality differences are shown by the heat map, as shown in Fig. 9.

For a single distribution of anomalies, such as a single abnormal color spot, a scratch, and a non-textured missing block, GMSD can better detect the difference between the

Table 4 ROC AUC indicator of representative algorithms on MVTec AD

Class	AE-SSIM [1]	AnoGAN [27]	RIAD [42]	DAGAN [32]	SCADN [38]	P-NET [44]	IGD [6]	FCDD [25]	Ours
Bottle	0.964	0.860	0.956	0.983	0.957	**0.990**	0.928	0.970	0.952
Cable	0.457	0.540	0.842	0.665	0.792	0.700	0.835	0.900	**0.940**
Capsule	0.780	0.840	0.928	0.687	0.765	0.840	**0.967**	0.930	0.964
Carpet	0.889	0.780	0.963	0.903	0.624	0.570	0.901	**0.960**	0.937
Grid	0.962	0.580	**0.988**	0.867	0.831	0.980	0.916	0.910	0.942
Hazelnut	0.827	0.830	0.961	**1.000**	0.856	0.970	0.981	0.950	0.976
Leather	0.871	0.640	**0.984**	0.944	0.983	0.890	0.983	0.980	0.960
Metal nut	0.677	0.760	0.925	0.815	0.504	0.790	0.902	**0.940**	0.911
Pill	0.759	0.870	0.957	0.768	0.833	0.910	0.962	0.810	**0.980**
Screw	0.681	0.800	0.988	**1.000**	0.968	**1.000**	0.960	0.860	0.915
Tile	0.964	0.500	0.891	0.961	0.814	**0.970**	0.727	0.910	0.963
Toothbrush	0.952	0.900	0.972	0.950	0.863	**0.990**	0.974	0.940	0.959
Transistor	0.530	0.800	0.877	0.794	0.981	0.820	0.843	0.880	**0.931**
Wood	0.797	0.620	0.858	0.979	0.659	**0.980**	0.827	0.880	0.902
Zipper	0.788	0.780	**0.975**	0.781	0.846	0.900	0.932	0.920	0.947
Avg	0.792	0.743	0.937	0.873	0.818	0.890	0.909	0.920	**0.945**

The bold values represent the best results of the different detection methods

Table 5 ROC AUC indicator of representative algorithms on MNIST

Model	0	1	2	3	4	5	6	7	8	9	Avg
AE	0.995	0.999	0.907	0.945	0.950	0.961	0.986	0.966	0.849	0.966	0.952
VAE [20]	0.997	0.999	0.935	0.958	0.973	0.962	0.991	0.975	0.921	0.976	0.966
AnoGAN [27]	0.966	0.992	0.850	0.887	0.894	0.883	0.947	0.935	0.849	0.924	0.913
ADGAN [8]	0.999	0.992	0.968	0.953	0.960	0.955	0.980	0.950	0.959	0.965	0.968
Ours	0.998	0.999	0.958	0.971	0.970	0.955	0.995	0.981	0.960	0.991	0.978

Table 6 Different optimization versions of the autoencoder model in terms of class-level testing

Model	MNIST
AE	0.952
VAE [20]	0.966
MemAE [12]	0.975
SCADN [38]	0.977
Ours	0.978

abnormal sample and the generated normal sample. It is precisely because GMSD locates the most prominent area. In the case where the foreground area is small and the background area is large and light-colored, errors can occur in the background localization. As in Fig. 10, SSIM and MSGMS can have large false detections, while GMSD focuses on highlighting the points with the largest differences in image quality, reducing the occurrence of false detections.

Due to the challenging nature of learning, blurred reconstructed images can still occur for object types where the normal sample shape is not fixed. In this case this paper uses GMSD and MSGMS as strategies for anomaly assessment. As shown in Fig. 11, when the reconstructed image appears blurred, this method is able to locate the area where the image quality difference is more prominent and locate the location of the anomaly more precisely. Even for normal images, when performing anomaly localization, it still goes for the closest anomaly. GMSD and MSGMS will produce fewer false detections, while SSIM will produce more obvious false detections. This is caused by blurred images with large texture gaps.

Experiments show that different settings of the proposed method will affect the overall performance of the model. By choosing the most appropriate method with prior knowledge of different data sets, the results can be further improved.

4.4 Experimental results

In this paper, Delete Paste, MFFE, and IQA modules are used for anomaly detection experiments, and the heatmap method is used to locate. The positioning results are shown in Fig. 6, which are visually satisfactory.

When testing, this paper uses Gaussian smoothing with a step size of 21 to extract abnormal scores. The results of the ROC AUC evaluation for localization at the pixel level are reported in Table 4, listing the results of anomalous localization on the MVTec AD data set by AE-SSIM, AnoGAN, RIAD, DAGAN, SCADN, P-NET, IGD, FCDD and the methods in this paper. The AUC of the RAMFAE method in this paper reaches 94.5. The method in this paper is superior to the advanced IGD, FCDD and RIAD, and 3.6, 2.5 and 0.8 higher than their detection methods.

The MNIST dataset has classes from ‘0’ to ‘9’ and is the most widely used dataset for one-class anomaly detection. The proposed algorithm is slightly better than other methods using this dataset. We modify the network parameters according to the MNIST dataset, and change the maximum number of channels of EncoderBlock, MFFE, and DecoderBlock to 128. MNIST is processed as a grayscale image. The proposed method is superior to the compared methods. The images in MNIST only contain simple patterns and can be well modeled by simply using MSE. The results of comparing the proposed model with general anomaly detection methods are shown in Table 5. Compared with AE, VAE, AnoGAN and ADGAN, our method is better. The best results were obtained on numbers ‘3’, ‘6’, ‘7’, ‘8’, ‘9’, with an average ROC AUC score of 97.8%. As Table 6 shows, RAMFAE achieves better performance in an autoencoder model with a similar capacity. The multi-scale feature extraction of MFFE plays a good role in image reconstruction. Our method is 1.2%, 0.3% and 0.1% higher than VAE, MemAE and SCADN.

5 Conclusion

This paper presents a novel visual anomaly detection method RAMFAE to solve the problem of image anomaly detection. Delete Paste is a novel data augmentation strategy for generating two different types of random exceptions. The cut part is pasted to a random position, and the pixels in the original position are deleted to 0 to avoid model degradation and abnormal reconstruction. By the way, a Multi-Scale Feature Focused Extraction network structure is designed, which is used between the encoder and the decoder to solve the image blurring problem and make the model pay more attention to the normal area. In addition, the comprehensive Image Quality Assessment focuses on both texture and local key information, enabling more accurate image reconstruction. In this paper, anomaly assessment was implemented on the MVTec AD anomaly data set, and the final ROC AUC result reached 94.5. It can be known that the Delete Paste and the MFFE are helpful for anomaly detection. The method in this paper still has some shortcomings. This paper finds that the image reconstruction is very good, but there is still the phenomenon of inaccurate positioning. It is related to the difficulty of reproducing the brightness of the image and the large background area. Future work will consider applying more feature extraction techniques, attention mechanisms, and better Image Quality Assessment to RAMFAE.

Acknowledgements This work was supported by the National Nature Science Foundation (61503038, 61403042); Scientific research project of Education Department of Liaoning Province (LQ2020013, LJKMZ20221484); a grant from Bohai University Teaching Reform Program (No. YJG20210023); a grant from Ministry of Education industry-University Cooperative Education Program (202102599009, 202101332004, 202101337001, 220504643183656); Application Basic Research Plan of Liaoning Province (2022JH2/101300282);Liaoning Natural Science Foundation under Grant (No. 2023-MS-294).

Data availability The data are available from the corresponding author on reasonable request.

References

1. Bergmann P, Löwe S, Fauser M et al (2018) Improving unsupervised defect segmentation by applying structural similarity to autoencoders. arXiv:1807.02011
2. Bergmann P, Fauser M, Sattlegger D et al (2021) Mvtec ad-a comprehensive real-world dataset for unsupervised anomaly detection. Int J Comput Vis 129:1038–1059
3. Bochkovskiy A, Wang CY, Liao HYM (2020) Yolov4: optimal speed and accuracy of object detection. arXiv:2004.10934
4. Cao Y, Zhao N, Xu N et al (2022) Minimal-approximation-based adaptive event-triggered control of switched nonlinear systems with unknown control direction. Electronics 11:33–86
5. Chen P, Liu S, Zhao H et al (2020) Gridmask data augmentation. arXiv:2001.04086
6. Chen Y, Tian Y, Pang G et al (2022) Deep one-class classification via interpolated gaussian descriptor, vol 36, pp 383–392
7. Chung H, Park J, Keum J et al (2020) Unsupervised anomaly detection using style distillation. IEEE Access 8:221494–221502
8. Deecke L, Vandermeulen R, Ruff L et al (2018) Anomaly detection with generative adversarial networks, vol 11051, pp 3–17
9. DeVries T, Taylor GW (2017) Improved regularization of convolutional neural networks with cutout. arXiv:1708.04552
10. Glorot X, Bordes A, Bengio Y (2011) Deep sparse rectifier neural networks, vol 15, pp 315–323
11. Golan I, El-Yaniv R (2018) Deep anomaly detection using geometric transformations. arXiv:1805.10917
12. Gong D, Liu L, Le V et al (2019) Memorizing normality to detect anomaly: memory-augmented deep autoencoder for unsupervised anomaly detection. arXiv:1904.02639
13. Goodfellow I, Pouget-Abadie J, Mirza M et al (2020) Generative adversarial networks. Commun ACM 63:139–144
14. Hendrycks D, Mazeika M, Kadavath S et al (2019) Using self-supervised learning can improve model robustness and uncertainty. Advances in neural information processing systems, p 32
15. Hinton GE, Zemel R (1993) Autoencoders, minimum description length and Helmholtz free energy. Advances in neural information processing systems, p 6

16. Hou Q, Zhou D, Feng J (2021) Coordinate attention for efficient mobile network design. arXiv:2103.02907
17. Hu J, Shen L, Sun G (2017) Squeeze-and-excitation networks. IEEE Trans Pattern Anal Mach Intell 42:2011–2023
18. Ioffe S, Szegedy C (2015) Batch normalization: accelerating deep network training by reducing internal covariate shift, vol 37, pp 448–456
19. Kang X, Zhang X, Li S et al (2017) Hyperspectral anomaly detection with attribute and edge-preserving filters. IEEE Trans Geosci Remote Sens 55:5600–5611
20. Kingma DP, Welling M (2014) Auto-encoding variational Bayes. arXiv:1312.6114
21. Li CL, Sohn K, Yoon J et al (2021) Cutpaste: self-supervised learning for anomaly detection and localization. arXiv:2104.04015
22. Li H, Zhu F, Qiu J (2018) Cg-diqa: no-reference document image quality assessment based on character gradient. arXiv:1807.04047
23. Lin D, Cao Y, Zhu W et al (2020) Few-shot defect segmentation leveraging abundant normal training samples through normal background regularization and crop-and-paste operation. arXiv:2007.09438
24. Liu X, Yang L, Chen J et al (2022) Region-to-boundary deep learning model with multi-scale feature fusion for medical image segmentation. Biomed Signal Process Control 71:103–165
25. Liznerski P, Ruff L, Vandermeulen RA et al (2020) Explainable deep one-class classification. arXiv:2007.01760
26. Russakovsky O, Deng J, Su H et al (2015) Imagenet large scale visual recognition challenge. Int J Comput Vis 115:211–252
27. Schlegl T, Seeböck P, Waldstein SM et al (2017) Unsupervised anomaly detection with generative adversarial networks to guide marker discovery, pp 146–157. arXiv:1703.05921
28. Schlegl T, Seeböck P, Waldstein SM et al (2019) f-anogan: fast unsupervised anomaly detection with generative adversarial networks. Med Image Anal 54:30–44
29. Schlüter HM, Tan J, Hou B et al (2022) Natural synthetic anomalies for self-supervised anomaly detection and localization, pp 474–489. arXiv:2109.15222
30. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556
31. Tack J, Mo S, Jeong J et al (2020) Csi: novelty detection via contrastive learning on distributionally shifted instances. Adv Neural Inf Process Syst 33:11839–11852
32. Tang TW, Kuo WH, Lan JH et al (2020) Anomaly detection neural network with dual auto-encoders gan and its industrial inspection applications. Sensors 20:33–36
33. Tao X, Wang Z, Zhang Z et al (2018) Wire defect recognition of spring-wire socket using multitask convolutional neural networks. IEEE Trans Compon Packag Manuf Technol 8:689–698
34. Veit A, Wilber MJ, Belongie S (2016) Residual networks behave like ensembles of relatively shallow networks. Advances in neural information processing systems, p 29
35. Wang Z, Bovik AC, Sheikh HR et al (2004) Image quality assessment: from error visibility to structural similarity. IEEE Trans Image Process 13:600–612
36. Woo S, Park J, Lee JY et al (2018) Cbam: convolutional block attention module, vol 11211, pp 3–19
37. Xue W, Zhang L, Mou X et al (2013) Gradient magnitude similarity deviation: a highly efficient perceptual image quality index. IEEE Trans Image Process 23:684–695
38. Yan X, Zhang H, Xu X et al (2021) Learning semantic context from normal samples for unsupervised anomaly detection, vol 35, pp 3110–3118
39. Yang H, Zhou Q, Song K et al (2020) An anomaly feature-editing-based adversarial network for texture defect visual inspection. IEEE Trans Ind Inf 17:2220–2230
40. Yang J, Shi Y, Qi Z (2020b) Dfr: deep feature reconstruction for unsupervised anomaly segmentation. arXiv:2012.07122
41. Yang Z, Bozchalooi IS, Darve E (2020c) Regularized cycle consistent generative adversarial network for anomaly detection. arXiv:2001.06591
42. Zavrtanik V, Kristan M, Skočaj D (2021) Reconstruction by inpainting for visual anomaly detection. Pattern Recognit 112:107706–107722
43. Zhang H, Cisse M, Dauphin YN et al (2017) mixup: beyond empirical risk minimization. arXiv:1710.09412
44. Zhou K, Xiao Y, Yang J et al (2020) Encoding structure-texture relation with p-net for anomaly detection in retinal images, vol 12365, pp 360–377

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.