J. Vis. Commun. Image R. 97 (2023) 103983

Contents lists available at ScienceDirect

J. Vis. Commun. Image R.

journal homepage: www.elsevier.com/locate/jvci

Full length article

Unsupervised industrial anomaly detection with diffusion models ☆

Haohao Xu [a], Shuchang Xu [a,*], Wenzhen Yang [b]

[a] School of Information Science and Technology, HangZhou Normal University, China
[b] Intelligent Perception Research Institute, Zhejiang Lab, China

ARTICLE INFO

Keywords:
Anomaly detection
Diffusion model
Image reconstruction
Unsupervised learning

ABSTRACT

Due to the limitations of autoencoders and generative adversarial networks, the performance of reconstruction-based unsupervised image anomaly detection methods are not satisfactory. In this paper, we aim to explore the potential of a more powerful generative model, the diffusion model, in the anomaly detection problem. Specifically, we design a Reconstructed Diffusion Models (RecDMs) based on conditional denoising diffusion implicit models for image reconstruction. To eliminate the stochastic nature of the generation process, our key idea is to use a learnable encoder to extract meaningful semantic representations, which are then used as signal conditions in an iterative denoising process to guide the model in recovering the image, while avoiding falling into an "identical shortcut" to meaningless image reconstruction. To accurately locate anomaly regions, we introduce a discriminative network to obtain the pixel-level anomaly segmentation map based on the reconstructed image. Our experiments demonstrate the effectiveness of the proposed method, achieving a new state-of-the-art image-level AUC score of 98.1% and a pixel-level AUC score of 94.6% on the MVTec AD dataset, among all reconstruction-based methods. We also show the significant potential and promising future of our method on the challenging real-world dataset, the CHL AD dataset.

1. Introduction

Image anomaly detection is often defined as a method of utilizing computer vision and machine learning techniques to detect anomalous regions within an image. Anomalous regions typically refer to areas within the image that differ from or do not conform to expected norms when compared to other regions, such as defects, stains, damages etc (see Fig. 1). The primary objective of image anomaly detection is to determine whether an image is anomalous, and furthermore, to indicate the anomalous regions. This technology can be applied to various application fields, including industrial product quality inspection, medical disease diagnosis [1] etc.

In recent years, anomaly detection methods based on deep learning have become one of the hottest research topics. In practical applications, due to the difficulty of obtaining anomaly samples and the impossibility of covering all anomaly patterns, it is usually assumed that only normal samples are available for model training, and test dataset containing few anomaly samples are used to verify the model's effectiveness. Therefore, anomaly detection actually can be treated as one-class classification problems, and typically employs unsupervised learning or semi-supervised learning techniques. For example, Lukas Ruff et al. [2] proposed Deep SVDD, which trains an encoder that attempts to map normal samples onto a hyper-spherical distribution, with the objective of minimizing the volume of the hyper-sphere. However, this method is unable to locate the anomalous regions, and the algorithm lacks of interpretability. Similarly, by learning the normal data distribution, the reconstruction-based method trains a model (AE [3], GANs [4–7]) to reconstruct input images. However, these methods still have some issues, such as the potential blurred and distorted images generated by autoencoders and the possibility of mode collapse and unstable training with GANs [8,9]. Additionally, due to the powerful feature extraction capabilities of convolutional neural networks [10,11] and the significant differences between normal and anomalous image features, embedding-based methods have shown strong competitiveness in the field of image anomaly detection [12–14]. However, their effectiveness cannot be guaranteed when a specific task is concerned, since pre-trained models are not always adapted to the task-specific dataset.

Recently, a new generative model, the Diffusion Model and its conditional variants [15–17], have shown great potential in image generation capabilities and have achieved impressive results in fields such as text-to-image [18,19], image super-resolution [20,21], and image translation [22,23]. In this paper, we explore the possibility of using the diffusion model in the field of image anomaly detection. Specifically, we design a novel diffusion model framework for image reconstruction,

☆ This paper has been recommended for acceptance by Dr. Zicheng Liu.

* Corresponding author.
E-mail addresses: xusc@hznu.edu.cn (S. Xu), ywz@zhejianglab.edu.cn (W. Yang).

https://doi.org/10.1016/j.jvcir.2023.103983
Received 25 May 2023; Received in revised form 28 September 2023; Accepted 11 November 2023
Available online 15 November 2023
1047-3203/

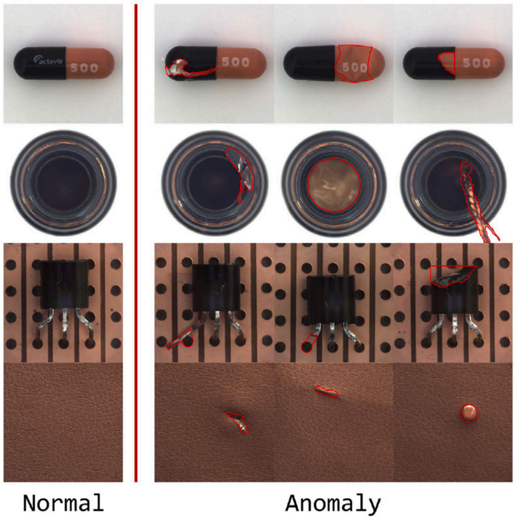


Fig. 1. Illustrations of image anomaly detection.

similar to the idea behind autoencoders, use a learnable encoder to extract meaningful semantic representations, which are then used as signal conditions in an iterative denoising process to guide the model in recovering the image while avoiding falling into an "identical shortcut" to meaningless image reconstruction. Clearly, after training on dataset containing only normal samples, the model can be reconstructed well for normal samples, while it will produce large reconstruction errors for abnormal samples that have never been seen. However, detail loss between reconstructed image and input image is still inevitable and significant noise will be introduced in following anomaly regions segmentation. To avoid possible noise produced in reconstruction step, we introduce a discriminative network to identify the differences between the input and reconstructed images. The end-to-end discriminative network, inspired by DRAEM [24] and self-supervised learning technology, can obtain accurate pixel-level segmentation map of anomaly regions. Furthermore, we use simulated anomaly samples to encourage the model to discover local irregularities in the anomaly samples. To validate the proposed method, we conducted experiments on the MVTec AD dataset [25], which includes 15 classes of industrial defect data in two categories: texture and object. We merged data from all classes to train an unified model. Compared to existing reconstruction-based methods, our model achieved state-of-the-art (SOTA) performance with an image-level AUROC score of 98.1% and a pixel-level AUROC score of 94.6%.

In summary, our work has the following main contributions:

(1) We present a novel reconstruction-based unsupervised anomaly detection framework, which applies a conditional diffusion model to learn the data distribution of normal patterns. Our experiments demonstrate that our method shows more powerful performance in unsupervised anomaly detection tasks.

(2) We utilize an end-to-end discriminative network to locate ano maly regions by comparing the input image and reconstructed image, and achieve more precise segmentation results.

(3) The proposed method do not need to train separated model for different domain, such as different category in a dataset, while keeps the performance of the model. This significantly improves efficiency.

(4) Our method has achieved state-of-the-art performance on a new and challenging defect detection dataset, demonstrating the promising ability of proposed method for complex application with high sample distribution variance.

2. Related work

In order to provide a clearer description of existing researches related to anomaly detection, we primarily divide the studies into three major categories: Embedding-Based, Anomaly Simulation-Based and Reconstruction-Based Methods as follows.

2.1. Embedding-based methods

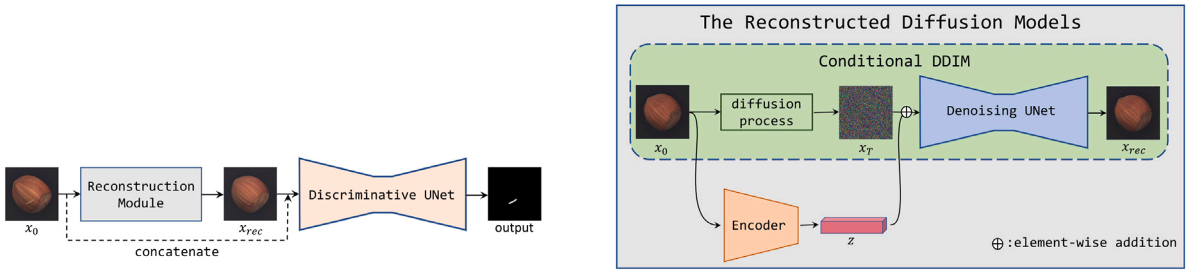
The image features of normal and anomalous samples typically exhibit significant differences, and convolutional neural network models pre-trained on large external datasets, such as ImageNet [26], have been demonstrated to possess powerful image feature extraction capabilities [27]. This has been extensively applied to tasks that suffer from data scarcity or extreme imbalance [10]. Based on the aforementioned points, embedding-based image anomaly detection methods utilize a pre-trained model to extract the features of normal samples, which are then used to construct a feature library comprising all the normal sample features. During the testing phase, the identical model is employed to extract the attributes of the testing sample. Subsequently, machine learning algorithms such as K-nearest neighbors (KNN) and distance metric functions like the Mahalanobis distance [28] are utilized to compute the distance to the feature library. This distance then serves as the metric for determining the anomaly score. Moreover, to locate the anomalous regions, these methods first divide the image into overlapping patches, calculate the anomaly score for each patch, and then aggregate them to obtain an anomaly segmentation map. Recently, these methods [13,14] have achieved high AUC scores on the unsupervised anomaly detection dataset MVTec AD. However, due to the need to store the feature library of normal samples, which can be highly memory-consuming and result in slower inference speeds.

2.2. Anomaly simulation-based methods

As only normal samples are available as training data, Anomaly Simulation-Based Methods propose to simulate anomalies as a data augmentation technique to generate positive samples and train an end-to-end image anomaly detection model based on the principle of self-supervised learning. Specifically, CutPaste [29] cuts an image patch and pastes it at a random location of a larger image, and then trains a binary classification model based on the simulated anomalous samples. Its drawback is that it requires the use of the gradCAM [30] technique to obtain the heatmap of anomalous regions. DRAEM [24] adds additional texture images as noise to normal images to generate anomalous regions. This type of data augmentation method aims to simulate texture anomalies. Although the Anomaly Simulation-Based Methods aim to encourage the model to detect local irregularities based on the simulated anomalous samples, in practice, the model assigns higher anomaly scores to samples that are more similar to the simulated anomalous samples, which requires the simulated anomalous samples to be closer to the real distribution.

2.3. Reconstruction-based methods

Reconstruction-Based Methods learn the distribution of normal samples by training generative models to reconstruct input images, and since the model produces larger reconstruction errors for anomalous regions due to the difference in data distribution, the reconstruction error can serve as the anomaly score. An et al. [3] first encode the image into a low-dimensional manifold embedding using VAE, and then reconstruct the input image from the low-dimensional representation, where the reconstruction probability is used to determine whether it is anomalous. Bergmann et al. [31] propose to use a structural similarity-based perceptual loss function to examine the interdependence between local image regions, while considering brightness, contrast, and structural information, rather than simply comparing individual pixel values. For



(a) (b)

Fig. 2. (a) An overview of our framework. By feeding the input image x_0 into the reconstruction module, we obtain the reconstructed image x_{rec}. Next, x_0 and x_{rec} are channel-wise concatenated and fed as input to the discrimination network, which outputs the anomaly segmentation map. (b) Reconstruction Module. The Reconstructed Diffusion Models consists of a conditional DDIM and a learnable encoder. The input image x_0 is first passed through the encoder to obtain the vector z, which is then used as a condition to guide the reconstruction of x_0 using the conditional DDIM. Where, x_T is obtained by diffusing x_0 through a diffusion process.

GAN-based methods, Xia et al. [6] propose a self-supervised learning framework that combines generative and discriminative methods to enhance the representation learning ability of the generator. Liang et al. [7] propose a novel OCR-GAN network that processes the anomaly detection task from a frequency perspective and achieves excellent image-level AUC scores, its drawback is that it cannot locate anomalous regions.

Diffusion models, as a new type of generative model, have been applied to medical image anomaly detection. Based on DDPM [15], AnoDDPM [32] proposes a novel local diffusion anomaly detection strategy that can be scaled to high-resolution images and uses a new noise source instead of Gaussian noise during the diffusion process. [33] uses a classifier-guided diffusion model [17] as a weakly supervised anomaly detection method, at the cost of training a classifier. However, in practice, due to the randomness of the generation process of diffusion models, the current diffusion-based methods cannot reconstruct input images well. In this study, inspired by [19,34], we use a classifier-free conditional diffusion model [35] for accurate image reconstruction and complete the industrial image anomaly detection task based on the reconstruction idea.

3. Proposed method

In this section, we present an unsupervised image anomaly detection framework based on the Reconstructed Diffusion Models. An overview of our framework is shown in Fig. 2 (a). Similar to typical reconstruction-based methods, our two-stage framework constructs an image anomaly detector. In the first stage, the input image is accurately reconstructed (see Section 3.1). The second stage performs anomaly detection and localization based on the input image and the reconstructed image (see Section 3.2). We introduced a discriminative network to obtain the anomaly segmentation map and trained it end-to-end in a self-supervised manner using simulated anomaly samples.

3.1. The reconstructed diffusion models

The diffusion model consists of a forward process that adds Gaussian noise $\mathcal{N}$, defined as a Markov chain in DDPM [15], and a backward process that iteratively removes the noise. In the backward process, a neural network is trained to predict the noise, indirectly learning the data distribution. Afterwards, the diffusion model can transform samples from a standard Gaussian noise to samples from the empirical data distribution through an iterative denoising process, which can be used for image generation. To control the generated image, the conditional diffusion model can introduce an input signal to guide the denoising process. Following Diffusion Autoencoders(Diff-AE) [34], we use DDIM [16] to accelerate the sampling process.

Inspired by autoencoders, we first use an encoder to extract meaningful semantic representation z of x_0, which avoids falling into an "identical shortcut" to meaningless image reconstruction. The semantic representation z is then used as a conditional input to guide the diffusion model for reconstructing x_0 (see Fig. 2 (b)). Unlike Diff-AE, we do not use DDIM reverse sample to get the stochastic code as x_T, since it will enable Diff-AE to almost perfectly reconstruct the input, regardless of whether the data is in the specified distribution or not. Specifically, our RecDMs models $p_\theta(x_{t-1}|x_t, z)$ to approximate the inference distribution $q(x_{t-1}|x_t, x_0)$:

$$q(x_{t-1}|x_t, x_0) = \mathcal{N}(\sqrt{\alpha_{t-1}}x_0 \quad + \sqrt{1 - x_{t-1} - \sigma_t^2} \cdot \frac{x_t - \sqrt{\alpha_t}x_0}{\sqrt{1-\alpha_t}}, \sigma_t^2 I) \tag{1}$$

$$p_\theta(x_{t-1}|x_t, z) = \begin{cases} \mathcal{N}(f_\theta(x_t, 1, z), \sigma_1^2 I), & if\ t = 1 \\ q(x_{t-1}|x_t, f_\theta(x_t, t, z)), & otherwise \end{cases} \tag{2}$$

Where α_t and σ_t are hyperparameters, which are used to represent the noise level and control the image generation process respectively. $x_1, x_2, \ldots, x_T$ are latent variables in the same sample space as x_0. and T is set to some large number, e.g. 1,000. The hyperparameter settings and the definition of p_θ are kept consistent with DDIM [15].

The function f_θ is defined as the model prediction for x_0 and can be further represented by a noise prediction network $\epsilon_\theta(x_t, t, z)$:

$$f_\theta(x_t, t, z) = \frac{1}{\sqrt{\alpha_t}}(x_t - \sqrt{1-\alpha_t}\epsilon_\theta(x_t, t, z)) \tag{3}$$

We optimize the model parameters with the following loss function L_{simple} [15]:

$$L_{simple} = \sum_{t=1}^{T} \mathbb{E}_{x_0,\epsilon_t}[\|\epsilon_\theta(x_t, t, z) - \epsilon_t\|_2^2] \tag{4}$$

Where $\epsilon_t \in \mathbb{R}^{3\times h\times w} \sim \mathcal{N}(0, I), x_t = \sqrt{\alpha_t}x_0 + \sqrt{1-\alpha_t}\epsilon_t$. Note that this simplified loss function has been shown to optimize both DDPM [15] and DDIM [16], though not the actual variational lower bound.

Model Architecture: The encoder consists of the first half of a U-Net [36] architecture with attention mechanisms, and a 512-dimensio nal vector z is obtained by adaptive global pooling at the end. The conditional DDIM architecture is based on the class-conditional U-Net model [17], while we use z instead of class information as condition.

3.2. Discriminative network and anomaly simulation strategy

In reconstruction-based anomaly detection methods, the anomaly map is obtained by comparing the original image with its reconstruction using similarity functions such as SSIM [37]. Due to the bottleneck constraints, the images reconstructed by our RecDMs may lose some details, which do not affect the image-level anomaly discrimination,

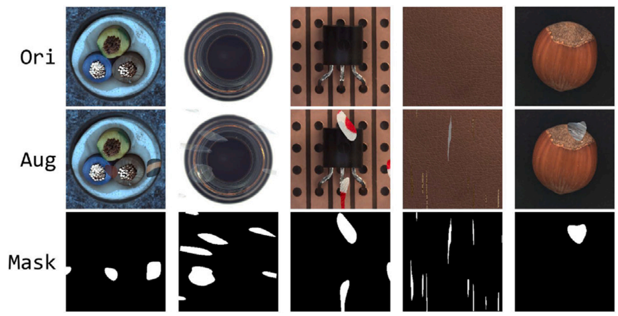


Fig. 3. Simulated anomaly results. Ori: Original images. Aug: augmented images. Mask: anomaly segmentation ground truth.

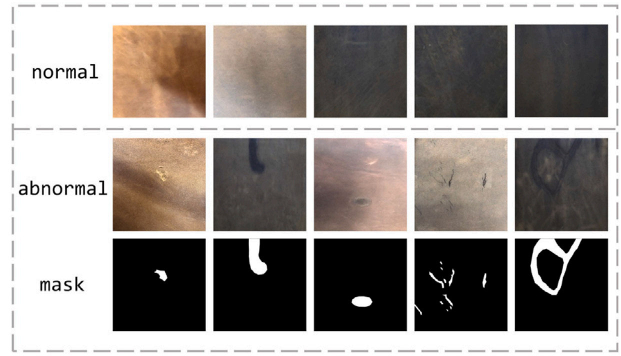


Fig. 4. Some examples in the CHL AD dataset.

but can easily affect handcrafted similarity functions and lead to noisy anomaly segmentation map, especially for normal samples. Therefore, we design a end-to-end discriminative network to learn an adaptive distance metric function between input image and reconstructed image. The network is then used to obtain a noise-free anomaly segmentation map.

Our discriminative network uses a U-Net architecture. The input of the network is defined as the channel-wise concatenation of the output of the conditional DDIM x_{rec} and the input image x_0. The network output is an anomaly segmentation map with the same size as the input image x_0. To address the issues of imbalanced positive and negative samples, we define the training objective of the discriminative network as Focal Loss [38], which is shown below:

$$FocalLoss(p,y) = \begin{cases} -\alpha(1-p)^{\gamma} log(p), & if\ y = 1 \\ -(1-\alpha)p^{\gamma} log(1-p), & if\ y = 0 \end{cases} \tag{5}$$

Where $y \in \{0,1\}$ specifies the ground-truth class, $p \in [0,1]$ is the model's estimated probability for the class with label $y = 1$, $\alpha \in [0,1]$ is a weighting factor to balance positive and negative samples, $\gamma \geq 0$ is a tunable focusing parameter to balance hard and easy samples.

The output of the discriminative network can be directly used as a pixel-level anomaly segmentation map. Similar to other anomaly detection methods [3,7,39,40], we first apply Gaussian mean filter to smooth the anomaly segmentation map and eliminate noise, and then take the highest pixel value of that as the image-level anomaly score.

For the anomaly simulation strategy, our goal is to generate appearances outside the distribution of normal samples, rather than simulating real anomaly samples. This encourage the discriminative network to learn a distance metric function that is specific to the current task based on local irregularities. Following DRAEM [24] and MemSeg [41], the noise images with various anomaly shapes are generated by Perlin noise generator [42], and then thresholded by random uniform sampling to obtain anomaly maps. The anomaly texture source images are sampled from an anomaly source image dataset, DTD [43], which is independent of the input image distribution. Fig. 3 shows some simulated anomaly results. Since our discriminative network automatically learns a suitable metric function based on image comparison, we have lower requirements for the fidelity of simulated anomalies, compared with anomaly simulation-based methods [24,41].

4. Experiment

4.1. Experimental setup

(1) Datasets.

The MVTec AD dataset [25] is a widely-used benchmark for semi-supervised and unsupervised image anomaly detection tasks, consisting of 5,354 high-resolution color images depicting 10 types of objects and 5 types of textures. These images have a resolution ranging from 700 to 1,024, and are downsampled to 256 × 256 resolution for all our experiments. Each object and texture category includes approximately 60 to 400 normal samples for training, along with a mixture of normal and anomaly images for testing. The test set contains a diverse range of realistic anomalies with different textures and scales, making it a challenging and comprehensive benchmark for evaluating anomaly detection algorithms.

The Crazy Horse Leather (CHL) AD dataset [44] consists of 104 color images showing variant colors and textures of Crazy Horse leather, which is widely used in leather products. Due to the complex manufacturing process, the images feature intricate backgrounds and a high variance in appearance distribution, encompassing multiple defect types such as steel stamps/brands, scars (healed wounds, knife marks), and wormholes. It is a highly challenging defect detection dataset, which contain only normal images in the training set as well. Images in this dataset are downsampled to 256 × 256 as well. Some examples in the dataset are given in Fig. 4.

Table 1
The hyperparameters of our RecDMs.

Batch size	16
z-dim	512
β scheduler	Linear
Learning rate	1e−4
Optimizer	Adam(no weight decay)
DDPM_T	1000
Loss	MSE with noise prediction ϵ
Var	No

(2) Implementation Details.

The baseline diffusion model and our RecDMs are based on the Diff-AE model [34]. We use an Nvidia A100 GPU for all our experiments. The hyperparameter settings are consistent with those of Diff-AE, as shown in the Table 1. In order to increase the sampling speed without compromising reconstruction quality, DDIM sampling step is set to T=250. We trained a unified reconstruction model by merging the training sets of all categories in the MVTec AD dataset.

The discriminative model utilizes the classic U-Net architecture with an input channel of 6. Focal loss ($\gamma = 2$) is used as the loss function, with a learning rate of $1e-4$. In our experiments, a separate discriminative model is trained for each category in the MVTec AD dataset to reduce the impact caused by the differences between categories.

We divide the training phase into two stages. Firstly, we train the reconstruction module using only normal samples in the training set. Then, we freeze the parameters of reconstruction module and train the discriminative network using simulated anomaly samples. Our code is available at https://github.com/xhh12381/RecDMs-AD.

(3) Evaluation Metrics

Anomaly detection performance is commonly evaluated using The Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC) Curve metric [7,13,14,24,29,41]. Higher AUC scores indicate better anomaly detection performance and are used as a standard evaluation metric for anomaly detection.

Table 2
We compared our method with other existing methods on the MVTec AD dataset, evaluating their performance using ROC-AUC scores, presented with the format of (Image-level, Pixel-level). Like us, the methods in three to five columns are reconstruction-based, while the following two columns are anomaly simulation-Based and embedding-based methods. Bold represents optimal results.

	Items	Skip-GAN [39]	Puzzle-AE [40]	OCR-GAN [7]	DRAEM [24]	PatchCore [14]	Ours
Texture	Carpet	(70.9,-)	(65.7,-)	(**98.9**,-)	(97.0,95.5)	(98.7,**99.0**)	(94.8,96.1)
	Grid	(47.7,-)	(75.4,-)	(99.6,-)	(**99.9**,**99.7**)	(98.2,98.7)	(99.5,93.4)
	Leather	(60.9,-)	(72.9,-)	(97.1,-)	(**100**,98.6)	(**100**,99.3)	(**100**,99.6)
	Tile	(29.9,-)	(65.5,-)	(**92.2**,-)	(99.6,99.2)	(98.7,95.6)	(**99.8**,93.5)
	Wood	(19.9,-)	(89.5,-)	(95.8,-)	(99.1,**96.4**)	(99.2,95.0)	(**99.6**,94.7)
	Average	(45.86,-)	(73.8,-)	(96.6,-)	(**99.1**,**97.9**)	(99.0,97.5)	(98.7,95.5)
Object	Bottle	(85.2,-)	(94.2,-)	(99.6,-)	(99.2,**99.1**)	(**100**,98.6)	(98.8,94.3)
	Cable	(54.4,-)	(87.9,-)	(99.2,-)	(91.8,94.7)	(**99.5**,**98.4**)	(92.1,94.7)
	Capsule	(54.3,-)	(66.9,-)	(95.4,-)	(**98.5**,94.3)	(98.1,**98.8**)	(97.9,97.2)
	Hazelnut	(24.5,-)	(91.2,-)	(88.2,-)	(**100**,**99.7**)	(**100**,98.7)	(98.9,95.0)
	Metal nut	(81.4,-)	(66.3,-)	(98.7,-)	(98.7,**99.5**)	(**100**,98.4)	(96.6,92.7)
	Pill	(67.1,-)	(71.6,-)	(98.5,-)	(**98.9**,**97.6**)	(96.6,97.4)	(95.3,95.6)
	Screw	(87.9,-)	(57.8,-)	(**100.0**,-)	(93.9,97.6)	(98.1,**99.4**)	(99.8,93.9)
	Toothbrush	(58.6,-)	(97.8,-)	(98.2,-)	(100,98.1)	(**100**,**98.7**)	(96.3,97.2)
	Transistor	(84.5,-)	(86.0,-)	(94.9,-)	(93.1,90.9)	(**100**,**96.3**)	(99.5,86.4)
	Zipper	(76.1,-)	(75.7,-)	(97.6,-)	(**100**,**98.8**)	(99.4,**98.8**)	(98.8,88.5)
	Average	(67.4,-)	(79.5,-)	(97.0,-)	(97.4,97.0)	(**99.2**,**98.4**)	(97.4,93.6)
	All	(60.2,-)	(77.6,-)	(96.9,-)	(98.0,97.3)	(**99.1**,**98.1**)	(98.1,94.6)

Table 3
The comparison between our method and patchcore [14] on the CHL AD dataset in terms of ROC-AUC % with the format of (Image-level, Pixel-level).

Methods	Results
PatchCore [14]	(0.77,0.54)
Ours	**(0.81,0.60)**

4.2. Comparisons with existing methods

For the MVTec AD dataset, Table 2 shows the AUC scores of different methods. Our method achieves a remarkable image-level AUC score of 98.1%, which outperforms existing reconstruction-based methods. In addition, all the existing reconstruction-based methods are not capable of giving pixel-level AUC scores, whereas our method achieves a highly competitive score of 94.6% at the pixel level in all image anomaly detection methods. Although it may not be the current state-of-the-art among all methods, as stated in earlier sections, reconstruction-based methods hold better interpretability and wider applicability than embedding-based methods in image anomaly detection. Nevertheless, our method exhibits relatively lower AUC scores in certain categories, such as cable and carpet. We attribute this to the significant semantic differences between the simulated anomalous samples and the real anomalous samples.

For the Crazy Horse Leather(CHL) AD dataset, our method achieved higher image-level and pixel-level AUC scores than PatchCore [14] (see Table 3). This indicates that our reconstruction-based method has an advantage for some applications with high sample distribution variance.

Additionally, we compared the memory consumption between our method and the embedding-based methods (taking PatchCore [14] as an example) on datasets with different sample numbers and image resolutions. Fig. 5 shows that our method has a lower memory consumption and is more advantageous for applications that require higher resolution images.

4.3. Ablation study

(1) reconstruction ability

The reconstruction capability of the model is the key factor in the performance of the proposed method. The quality of image reconstruction determines the accuracy of detecting and precisely localizing anomalous regions. In Table 4, we evaluate the reconstruction quality of four methods: 1) our RecDMs, 2) Puzzle-AE [40], 3) Skip-GAN [39], 4) OCR-GAN [7]. All these models are trained and tested on the MVTec AD dataset. The evaluation metrics are SSIM [37] (↑), LPIPS [45] (↓), and MSE(↓).

Furthermore, we aim to demonstrate that our method is only capable of reconstructing normal samples, while producing significantly larger reconstruction discrepancies for anomalous samples. As shown in Fig. 6, we visualize the reconstructed images and the difference images between the reconstructed and original images to explore the reconstruction ability of different methods. The results clearly indicate that the images reconstructed using Puzzle-AE [40] exhibit significant distortion and blurring. Although OCR-GAN [7] is capable of generating high-quality reconstructions, its excessive model generalization

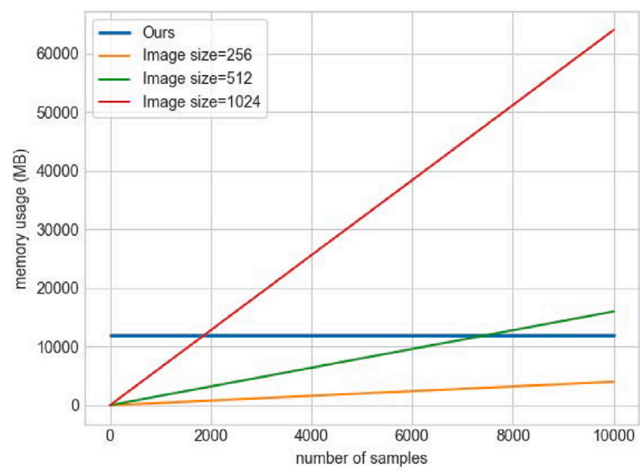


Fig. 5. Memory consumption comparisons between our model and PatchCore [12] on datasets with different sample numbers and image resolutions.

Table 4
The reconstruction quality comparison between our method and different methods on the MVTec AD dataset in terms of SSIM, LPIPS and MSE.

Models	SSIM (↑)	LPIPS (↓)	MSE (↓)
Puzzle-AE [40]	0.592	0.187	0.153
Skip-GAN [39]	0.645	0.208	0.171
OCR-GAN [7]	0.766	0.158	0.049
Ours	**0.857**	**0.141**	**0.012**

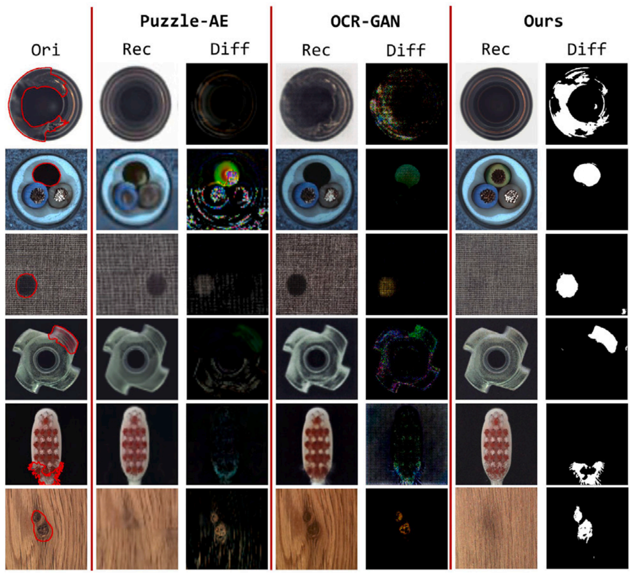


Fig. 6. Reconstruction results comparison of our method with other two reconstruction-based methods. Ori: Original images with anomaly segmentation ground truth. Rec: Reconstructed images. Diff: Differences between original and reconstructed images.

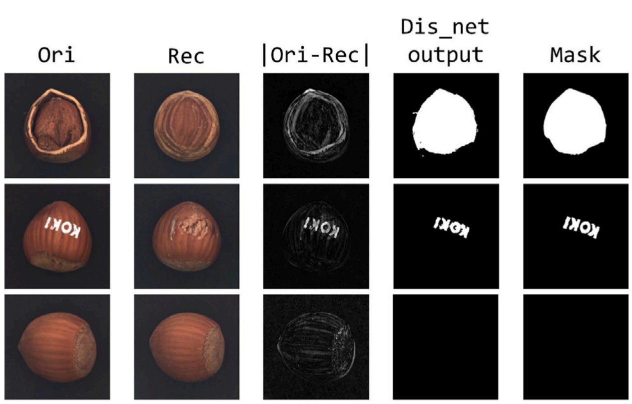


Fig. 7. The effect of the discriminative network on anomaly localization. Directly using |Ori-Rec | as an anomaly segmentation map results in additional noise, while using the discriminative network yields a more accurate segmentation map. Dis_net output: Discriminative network output. Mask: Ground Truth.

capability leads to effective reconstruction of anomaly regions that should not be accurately reconstructed. Reconstructed results indicate that our RecDMs exhibits superior capability in preserving fine details, while the difference images effectively highlight the anomaly regions, outperforming other classical reconstruction-based methods.

(2) The discriminative network effectiveness

The discriminative network is responsible for generating accurate and noise-free anomaly segmentation maps. In Fig. 7, we demonstrate the difference between having and not having discriminative network. It is evident that subtracting the reconstructed images directly from the original ones as the anomaly segmentation map results in significant noise due to the loss of some unnecessary details in the reconstruction process. However, the discriminative network effectively removes this noise.

5. Conclusion & limitations

In the field of unsupervised image anomaly detection, reconstruc tion-based methods are a classic technique that has perfect theoretical feasibility and good interpretability. However, their suboptimal reconstruction quality often leads to unsatisfactory performance. In this paper, we propose a novel reconstruction-based unsupervised image anomaly detection method by combining state-of-the-art generative models, specifically diffusion models, followed by an end-to-end discriminative network to generate precise anomaly segmentation map. Our experimental results demonstrate impressive performance, indicating that reconstruction-based methods can also exhibit robust capabilities.

Limitation 1: The inference speed of the diffusion model is hindered by the requirement for a large number of sampling steps, even with the use of DDIM to enhance sampling efficiency. As a result, our method exhibits subpar performance in terms of time efficiency. Limitation 2: The involvement of a discriminative network in generating simulated anomaly samples introduces additional training data and may lead to domain shift problems. In future work, we intend to explore improved post-processing methods, as this is a common challenge in reconstruction-based approaches.

In the future, our work aims to discover better post-processing methods, which is actually a common issue in the reconstruction-based methods.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This paper is supported by the National Key Research and Development Program of China (2021YFF0600203); the Major Scientific Research Project of the Zhejiang Laboratory (2022MG0AC04).

References

[1] X. Liu, Z. Liu, Y. Zhang, M. Wang, J. Tang, Weakly-supervised localization and classification of biomarkers in oct images with integrated reconstruction and attention, Biomed. Signal Process. Control 79 (2023) 104213.

[2] L. Ruff, R. Vandermeulen, N. Goernitz, L. Deecke, S.A. Siddiqui, A. Binder, E. Müller, M. Kloft, Deep one-class classification, in: International Conference on Machine Learning, PMLR, 2018, pp. 4393–4402.

[3] J. An, S. Cho, Variational autoencoder based anomaly detection using reconstruction probability, Special Lecture on IE 2 (1) (2015) 1–18.

[4] T. Schlegl, P. Seeböck, S.M. Waldstein, U. Schmidt-Erfurth, G. Langs, Unsupervised anomaly detection with generative adversarial networks to guide marker discovery, in: Information Processing in Medical Imaging: 25th International Conference, IPMI 2017, Boone, NC, USA, June 25-30, 2017, Proceedings, Springer, 2017, pp. 146–157.

[5] T. Schlegl, P. Seeböck, S.M. Waldstein, G. Langs, U. Schmidt-Erfurth, f-AnoGAN: Fast unsupervised anomaly detection with generative adversarial networks, Med. Image Anal. 54 (2019) 30–44.

[6] X. Xia, X. Pan, X. He, J. Zhang, N. Ding, L. Ma, Discriminative-generative representation learning for one-class anomaly detection, 2021, arXiv preprint arXiv:2107.12753.

[7] Y. Liang, J. Zhang, S. Zhao, R. Wu, Y. Liu, S. Pan, Omni-frequency channel-selection representations for unsupervised anomaly detection, 2022, arXiv preprint arXiv:2203.00259.

[8] M. Arjovsky, S. Chintala, L. Bottou, Wasserstein generative adversarial networks, in: International Conference on Machine Learning, PMLR, 2017, pp. 214–223.

[9] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, A.C. Courville, Improved training of wasserstein gans, in: Advances in Neural Information Processing Systems, Vol. 30, 2017.

[10] D. Hendrycks, K. Lee, M. Mazeika, Using pre-training can improve model robustness and uncertainty, in: International Conference on Machine Learning, PMLR, 2019, pp. 2712–2721.

[11] P. Perera, V.M. Patel, Learning deep features for one-class classification, IEEE Trans. Image Process. 28 (11) (2019) 5450–5463.

[12] O. Rippel, P. Mertens, D. Merhof, Modeling the distribution of normal data in pre-trained deep features for anomaly detection, in: 2020 25th International Conference on Pattern Recognition, ICPR, IEEE, 2021, pp. 6726–6733.

[13] T. Reiss, N. Cohen, L. Bergman, Y. Hoshen, Panda: Adapting pretrained features for anomaly detection and segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 2806–2814.

[14] K. Roth, L. Pemula, J. Zepeda, B. Schölkopf, T. Brox, P. Gehler, Towards total recall in industrial anomaly detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 14318–14328.
[15] J. Ho, A. Jain, P. Abbeel, Denoising diffusion probabilistic models, Adv. Neural Inf. Process. Syst. 33 (2020) 6840–6851.
[16] J. Song, C. Meng, S. Ermon, Denoising diffusion implicit models, in: International Conference on Learning Representations, 2020.
[17] P. Dhariwal, A. Nichol, Diffusion models beat GANs on image synthesis, Adv. Neural Inf. Process. Syst. 34 (2021) 8780–8794.
[18] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, B. Ommer, High-resolution image synthesis with latent diffusion models, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 10684–10695.
[19] L. Zhang, M. Agrawala, Adding conditional control to text-to-image diffusion models, 2023, arXiv preprint arXiv:2302.05543.
[20] Z. Kadkhodaie, E.P. Simoncelli, Solving linear inverse problems using the prior implicit in a denoiser, 2020, arXiv preprint arXiv:2007.13640.
[21] C. Saharia, J. Ho, W. Chan, T. Salimans, D.J. Fleet, M. Norouzi, Image super-resolution via iterative refinement, IEEE Trans. Pattern Anal. Mach. Intell. (2022).
[22] H. Sasaki, C.G. Willcocks, T.P. Breckon, Unit-ddpm: Unpaired image translation with denoising diffusion probabilistic models, 2021, arXiv preprint arXiv:2104.05358.
[23] C. Saharia, W. Chan, H. Chang, C. Lee, J. Ho, T. Salimans, D. Fleet, M. Norouzi, Palette: Image-to-image diffusion models, in: ACM SIGGRAPH 2022 Conference Proceedings, 2022, pp. 1–10.
[24] V. Zavrtanik, M. Kristan, D. Skočaj, Draem-a discriminatively trained reconstruction embedding for surface anomaly detection, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 8330–8339.
[25] P. Bergmann, M. Fauser, D. Sattlegger, C. Steger, MVTec AD–A comprehensive real-world dataset for unsupervised anomaly detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 9592–9600.
[26] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, Ieee, 2009, pp. 248–255.
[27] M. Huh, P. Agrawal, A.A. Efros, What makes ImageNet good for transfer learning? 2016, arXiv preprint arXiv:1608.08614.
[28] R. De Maesschalck, D. Jouan-Rimbaud, D.L. Massart, The mahalanobis distance, Chemometr. Intell. Lab. Syst. 50 (1) (2000) 1–18.
[29] C.-L. Li, K. Sohn, J. Yoon, T. Pfister, Cutpaste: Self-supervised learning for anomaly detection and localization, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 9664–9674.
[30] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 618–626.
[31] P. Bergmann, S. Löwe, M. Fauser, D. Sattlegger, C. Steger, Improving unsupervised defect segmentation by applying structural similarity to autoencoders, 2018, arXiv preprint arXiv:1807.02011.
[32] J. Wyatt, A. Leach, S.M. Schmon, C.G. Willcocks, Anoddpm: Anomaly detection with denoising diffusion probabilistic models using simplex noise, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 650–656.
[33] J. Wolleb, F. Bieder, R. Sandkühler, P.C. Cattin, Diffusion models for medical anomaly detection, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part VIII, Springer, 2022, pp. 35–45.
[34] K. Preechakul, N. Chatthee, S. Wizadwongsa, S. Suwajanakorn, Diffusion autoencoders: Toward a meaningful and decodable representation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 10619–10629.
[35] J. Ho, T. Salimans, Classifier-free diffusion guidance, 2022, arXiv preprint arXiv:2207.12598.
[36] O. Ronneberger, P. Fischer, T. Brox, U-Net: Convolutional networks for biomedical image segmentation, in: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18, Springer, 2015, pp. 234–241.
[37] Z. Wang, A.C. Bovik, H.R. Sheikh, E.P. Simoncelli, Image quality assessment: From error visibility to structural similarity, IEEE Trans. Image Process. 13 (4) (2004) 600–612.
[38] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2980–2988.
[39] S. Akçay, A. Atapour-Abarghouei, T.P. Breckon, Skip-ganomaly: Skip connected and adversarially trained encoder-decoder anomaly detection, in: 2019 International Joint Conference on Neural Networks, IJCNN, IEEE, 2019, pp. 1–8.
[40] M. Salehi, A. Eftekhar, N. Sadjadi, M.H. Rohban, H.R. Rabiee, Puzzle-ae: Novelty detection in images through solving puzzles, 2020, arXiv preprint arXiv:2008.12959.
[41] M. Yang, P. Wu, H. Feng, MemSeg: A semi-supervised method for image surface defect detection using differences and commonalities, Eng. Appl. Artif. Intell. 119 (2023) 105835.
[42] K. Perlin, An image synthesizer, ACM Siggraph Comput. Graph. 19 (3) (1985) 287–296.
[43] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, A. Vedaldi, Describing textures in the wild, in: 2014 IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 3606–3613, http://dx.doi.org/10.1109/CVPR.2014.461.
[44] H.X. Shuchang Xu, The Crazy Horse Leather (CHL) AD dataset, 2023, https://github.com/xhh12381/RecDMs-AD.
[45] R. Zhang, P. Isola, A.A. Efros, E. Shechtman, O. Wang, The unreasonable effectiveness of deep features as a perceptual metric, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 586–595.