

GLAD: Towards Better Reconstruction with Global and Local Adaptive Diffusion Models for Unsupervised Anomaly Detection

Hang Yao¹, Ming Liu^{1,2}(✉), Zhicun Yin¹,
Zifei Yan^{1,2}, Xiaopeng Hong¹, and Wangmeng Zuo^{1,2}

¹ Harbin Institute of Technology, Harbin, China
{yaohang_1, csmliu, cszcyin}@outlook.com, amber_1980@163.com,
hongxiaopeng@ieee.org, cswmzuo@gmail.com
² Pazhou Lab Huangpu, Guangzhou, China

Abstract. Diffusion models have shown superior performance on unsupervised anomaly detection tasks. Since trained with normal data only, diffusion models tend to reconstruct normal counterparts of test images with certain noises added. However, these methods treat all potential anomalies equally, which may cause two main problems. From the global perspective, the difficulty of reconstructing images with different anomalies is uneven. For example, adding back a missing element is harder than dealing with a scratch, thus requiring a larger number of denoising steps. Therefore, instead of utilizing the same setting for all samples, we propose to predict a particular denoising step for each sample by evaluating the difference between image contents and the priors extracted from diffusion models. From the local perspective, reconstructing abnormal regions differs from normal areas even in the same image. Theoretically, the diffusion model predicts a noise for each step, typically following a standard Gaussian distribution. However, due to the difference between the anomaly and its potential normal counterpart, the predicted noise in abnormal regions will inevitably deviate from the standard Gaussian distribution. To this end, we propose introducing synthetic abnormal samples in training to encourage the diffusion models to break through the limitation of standard Gaussian distribution, and a spatial-adaptive feature fusion scheme is utilized during inference. With the above modifications, we propose a global and local adaptive diffusion model (abbreviated to GLAD) for unsupervised anomaly detection, which introduces appealing flexibility and achieves anomaly-free reconstruction while retaining as much normal information as possible. Extensive experiments are conducted on three commonly used anomaly detection datasets (MVTec-AD, MPDD, and VisA) and a printed circuit board dataset (PCB-Bank) we integrated, showing the effectiveness of the proposed method. The source code and pre-trained models are publicly available at <https://github.com/hyao1/GLAD>.

Keywords: Unsupervised Anomaly Detection · Diffusion Models · Adaptive Denoising Process

1 Introduction

Anomaly detection (AD) aims to detect and locate abnormal patterns that influence the appearance and function of objects, which is vital for the quality of products and has been widely used in industries [5, 16, 33]. In practice, the prevalence of different anomaly types varies, making it challenging to collect enough abnormal samples for all anomaly types in situations with high yield rates. Furthermore, due to the ever-changing product design and production processes, it is impossible to collect all anomalies in advance. Therefore, unsupervised anomaly detection (UAD) has drawn much attention with only normal samples required. To achieve unsupervised anomaly detection, reconstruction-based methods generate a potential normal sample corresponding to the given one, and the anomalies can be detected and located via the comparison between the given sample and its normal counterpart. Due to the prominent modeling ability, diffusion models are introduced for sample reconstruction and have shown superior performance.

Existing diffusion model-based UAD methods [9, 17, 18, 27] typically follow a common process to reconstruct the test samples. To begin with, a diffusion model is trained with normal samples of certain objects or products (*e.g.*, bottles, hazelnuts, *etc.*). Then, it can be deployed to reconstruct a sample with random noise added. Note that during the training process, the diffusion model captures the distribution of normal samples only, which implies that it will generate a normal sample from any noise-contaminated inputs as long as the randomness is strong enough³. Therefore, existing methods choose to set a sufficiently large denoising step to guarantee the reconstruction ability.

However, setting the same denoising step for all samples is a sub-optimal solution. As shown in Fig. 1, the difficulty of reconstructing images with different anomalies is uneven. For example, 900 steps are required to add a missing element back, while 300 steps are already enough to deal with a scratch. Besides, apart from better reconstruction ability, a larger denoising step also means higher randomness and uncertainty, leading to less preserved details of the original test samples (see the areas bounded by the red lines in Fig. 1.) To this end, we propose to set an Adaptive Denoising Step (ADP) for each sample, which achieves a better trade-off between reconstruction quality and detail preservation ability. In order to implement such an adaptive denoising step method, we take advantage of the prior in the diffusion models. Specifically, we first add noise to the test sample with a large enough noise weight, and perform the denoising steps to gradually remove the noises and reconstruct a normal sample. During the reconstruction procedure, we can compare the reconstructed sample with the noise-contaminated input, where the difference reflects the existence of anomalies and can help adaptively determine the denoising steps. Since this proper steps

³ In the setting of diffusion models, the randomness is equivalent to the weight of the random noise, which is determined by the denoising step. In other words, a larger denoising step means higher noise weight and stronger randomness.

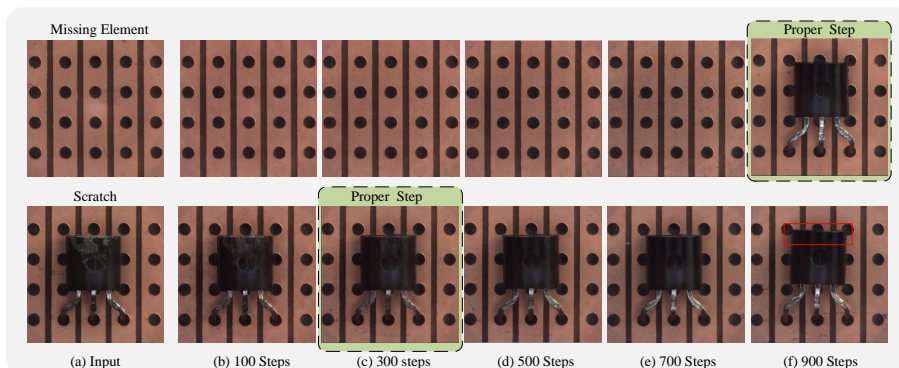


Fig. 1: Illustration of adaptive denoising process. For severe anomalies like missing elements, it requires a large number of denoising steps (900) to add the element back, while for small anomalies like scratch, 300 steps are already enough. Besides, setting a large enough denoising step (*e.g.*, 900) for all samples will affect the detail preservation. For example, in the area bounded by red lines, the position of the element is changed, which will be marked as anomalies during the comparison process.

is used to add noise to the whole image, we can regard it as a global adaptive setting.

Apart from the global denoising step, from a local perspective, we also find that even in the same image, reconstructing the abnormal regions is different from the normal areas. For normal areas, the diffusion model only needs to remove the added noise. It means that the noise to be predicted is exactly the one added, both following the standard Gaussian distribution. However, for abnormal regions, in order to reconstruct their normal counterparts, the noise to be predicted by the diffusion model inevitably deviates from the standard Gaussian distribution, making the prediction more difficult than in the normal areas. As a remedy, we propose Anomaly-oriented Training Paradigm (ATP), which introduces synthetic abnormal samples when training the diffusion models and generalizes the loss function of the diffusion model to a more general form. The proposed Anomaly-Oriented Training Paradigm encourages the diffusion models to break through the limitation of standard Gaussian distribution and promotes the ability to generate normal samples in the abnormal regions. Besides, for better detail preservation, we introduce an Spatial-Adaptive Feature Fusion (SAFF) scheme during inference by fusing sample features of the normal regions and generated features in abnormal regions, which better preserves the details in potential normal regions and alleviates the difficulty of the subsequent comparison procedure. In this way, the diffusion model becomes flexible enough to achieve a local adaptive inference for different regions (normal and abnormal regions), with both reconstruction and detail preservation abilities equipped.

With the above two modifications, a global and local adaptive diffusion model (abbreviated to GLAD) is presented. Extensive experiments are conducted on four anomaly detection datasets (*i.e.*, MVTec-AD [1], MPDD [11], VisA [33],

and the PCB-Bank dataset⁴) to verify the effectiveness of the proposed method. The experimental results show that our method achieves superior performance on unsupervised anomaly detection tasks. The contributions of this paper are as follows.

- For a better trade-off between reconstruction quality and detail preservation, unlike existing diffusion model-based methods utilizing the same setting for all samples, we propose to predict an Adaptive Denoising Step (ADP) as a global adaptive setting for each sample to retain more normal information.
- Considering the difference between abnormal regions and normal areas, we introduce Anomaly-oriented Training Paradigm (ATP) during training to allow diffusion model to predict non-Gaussian noise at abnormal regions, and propose a Spatial-Adaptive Feature Fusion (SAFF) scheme during inference to avoid reconstruction of abnormal regions.
- The experiments on three commonly used datasets and our integrated PCB-Bank show that the proposed global and local adaptive diffusion model (GLAD) improves both reconstruction quality and anomaly detection ability.

2 Related Work

2.1 Anomaly Detection

Mainstream unsupervised anomaly detection methods can be divided into two categories, *i.e.* reconstruction-based methods and embedding-based methods.

Reconstruction-based methods suppose that the model trained on normal samples can only reconstruct normal images well, not abnormal areas. Anomalies can be detected by comparing the samples before and after reconstruction. Early methods [3, 14, 32] utilize variational auto-encoders [12] to reconstruct samples. OCR-GAN [13] decouples images into different frequency components and models the reconstruction process as a combination of parallel omni-frequency image restorations. DRAEM [28] synthesizes pseudo-anomaly images to train a UNet for reconstruction. Recently, diffusion models [10, 24] are proposed and achieve state-of-the-art performance. DiffAD [29] diffuses the test sample as noisy condition to produce high-quality reconstructed images while retaining normal information. DDAD [18] uses score-based function to reintegrate the information of test samples during the denoising process. However, these methods add fixed steps of noise for denoising, which is not suitable for various anomaly types.

Embedding-based methods extract feature of images to evaluate abnormal areas. Knowledge distillation-based methods [2, 21] first train student network with normal samples. Then, features from the pre-trained teacher network are compared with features from the student network to detect and locate anomalies. Reverse distillation [7] is developed to utilize different architectures of teacher

⁴ PCB-Bank is a printed circuit board dataset we integrated from existing datasets, please refer to <https://github.com/SSRheart/industrial-anomaly-detection-dataset> for more details.

and student to maintain the distinction of anomaly. PaDiM [6] builds multivariate Gaussian distributions for patch features of normal samples and uses Mahalanobis distance as the anomaly score. PatchCore [20] proposes a memory bank to save features of normal images, which are compared with feature maps of test images to distinguish the difference between normal and abnormal features.

2.2 Diffusion Model

Inspired by principles of nonequilibrium thermodynamics [23], diffusion model (DM) [10] is proposed for image generation, and utilized in a variety of downstream tasks [15,25,30,31]. Denoising diffusion implicit models (DDIM) [24] considers the reverse process of DM as non-Markovian processes, which speeds up the inference greatly. Latent diffusion model (LDM) [19] conducts training and inference in latent space, further reducing the cost of resources and time. Besides, Text inversion [8] and Dreambooth [22] learn the appearance of subjects in a given reference set and synthesize novel renditions of them in different contexts. These methods follow the training paradigm for predicting Gaussian noise and start denoising from a Gaussian noise. Thus, these methods can not achieve adaptive denoising.

3 Methodology

In this section, we start with the common practice of existing diffusion model-based unsupervised anomaly detection methods. By transforming their working processes into formal expressions of formulas, we naturally reveal the existing problems and discover the corresponding solutions, which derive the global and local adaptive diffusion model (*i.e.*, GLAD) in this paper.

3.1 Preliminary

First, we provide preliminary knowledge of diffusion models for later analyses.

Diffusion Process. In the diffusion process, a random noise ϵ is added to the sample \mathbf{x} , and the result after t steps can be represented by,

$$\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon, \quad (1)$$

where $\bar{\alpha}_t$ is manually defined, which is negatively correlated with t , and $\mathbf{x}_0 = \mathbf{x}$.

Intermediate Result Visualization. Eq. (1) can be rewritten to obtain the noise-free version of the intermediate result at the t -th step, *i.e.*,

$$\mathbf{x}_{t \rightarrow 0} = \frac{1}{\sqrt{\bar{\alpha}_t}}(\mathbf{x}_t - \sqrt{1 - \bar{\alpha}_t}\epsilon_\theta(\mathbf{x}_t, t)), \quad (2)$$

where ϵ_θ is from the pre-trained diffusion model for predicting the noise added.

Generation Process. Each step of the generation stage can be formulated by,

$$\hat{\mathbf{x}}_{t-1} = \sqrt{\bar{\alpha}_{t-1}}\hat{\mathbf{x}}_{t \rightarrow 0} + \sqrt{1 - \bar{\alpha}_{t-1}}\epsilon_\theta(\hat{\mathbf{x}}_t, t). \quad (3)$$

Note that the variables predicted by the diffusion model are marked by \wedge , for example, \mathbf{x}_t is obtained by adding random noise directly (the diffusion process), while $\hat{\mathbf{x}}_t$ is obtained by denoising from a larger step (the generation process).

3.2 Formulaic Analysis on Reconstruction Errors.

In existing diffusion model-based unsupervised anomaly detection methods, the common way of reconstructing the normal counterpart of a test sample is to add certain noise to the given test sample and then execute the generation process of the diffusion model. Denote the test sample with anomalies by \mathbf{x}^a , its potential normal counterpart by \mathbf{x} , then the process can be described by $\mathbf{x}^a \xrightarrow{\text{diff}} \mathbf{x}_t^a \xrightarrow{\text{gen}} \hat{\mathbf{x}}_t^a$, and ideally we should have $\hat{\mathbf{x}}_t^a \rightarrow \mathbf{x}$. Since the anomalies are typically detected and located by comparison between $\hat{\mathbf{x}}_t^a$ and \mathbf{x}^a , we can require that $\|\hat{\mathbf{x}}_t^a - \mathbf{x}\|_\infty < \tau$, where τ is a threshold manually set for distinguishing between normal and abnormal samples.

Since the diffusion model is pre-trained and fixed during the generation process, we analyze in the t step. Denote the difference between \mathbf{x}^a and \mathbf{x} by \mathbf{n} , *i.e.*, $\mathbf{x}^a = \mathbf{x} + \mathbf{n}$, and according to Eq. (1),

$$\begin{aligned} \mathbf{x}_t^a &= \sqrt{\bar{\alpha}_t} \mathbf{x}_0^a + \sqrt{1 - \bar{\alpha}_t} \epsilon^a \\ &= \sqrt{\bar{\alpha}_t} \mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon^a + \sqrt{\bar{\alpha}_t} \mathbf{n}, \end{aligned} \quad (4)$$

where ϵ^a is the noise added to \mathbf{x}^a . Denote the generation process from step t by g_t , by combining Eqs. (1) and (4), the error can be represented by,

$$\begin{aligned} \hat{\mathbf{x}}_t^a - \mathbf{x} &= g_t(\mathbf{x}_t^a) - g_t(\mathbf{x}_t) \\ &= g_t(\sqrt{\bar{\alpha}_t} \mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon^a + \sqrt{\bar{\alpha}_t} \mathbf{n}) - g_t(\sqrt{\bar{\alpha}_t} \mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon) \\ &\stackrel{\approx}{\sim} \sqrt{1 - \bar{\alpha}_t} (\epsilon^a - \epsilon) + \sqrt{\bar{\alpha}_t} \mathbf{n}, \end{aligned} \quad (5)$$

where $\stackrel{\approx}{\sim}$ means approximately proportional to, which is based on a reasonable assumption that g_t is smooth enough. In the following, we make our efforts to reduce the errors in Eq. (5) for a better reconstruction quality.

3.3 Adaptive Denoising Steps

In existing methods, since the noise is always assumed to follow a standard Gaussian distribution, the error between ϵ^a and ϵ are ignored, and then only $\sqrt{\bar{\alpha}_t} \mathbf{n}$ leaves in Eq. (5). Considering the requirement $\|\hat{\mathbf{x}}_t^a - \mathbf{x}\|_\infty < \tau$, for larger \mathbf{n} , a smaller $\sqrt{\bar{\alpha}_t}$ is desired. Since $\bar{\alpha}_t$ is negatively correlated with t , a smaller $\sqrt{\bar{\alpha}_t}$ means a larger t , which provides a formulaic explanation for our motivation to set a proper denoising step for each sample.

With the above analysis, an intuitive way to determine the proper step is by evaluating the value of \mathbf{n} , which, however, is a concept we introduced and is unavailable in practice. Actually, in the generation process, \mathbf{n} is reflected in $\epsilon_\theta^a = \epsilon_\theta(\mathbf{x}_t^a, t)$, which can be used for comparison.

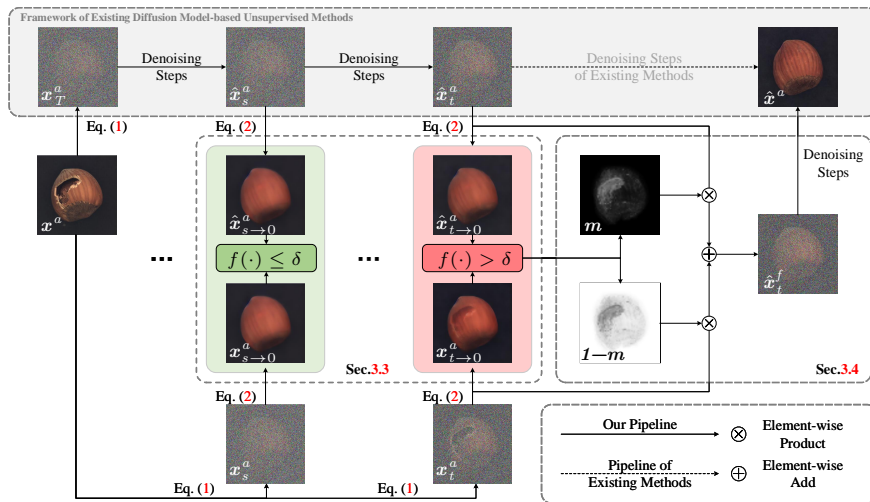


Fig. 2: The reconstruction pipeline of the proposed GLAD, including the Adaptive Denoising Steps (Sec. 3.3) and the Spatial-Adaptive Feature Fusion Scheme (Sec. 3.4).

As shown in the left part of Fig. 2, starting from the input sample \mathbf{x}^a , we add sufficient noises (e.g., T steps) and obtain \mathbf{x}_T^a . Then, in each step, we compare \mathbf{x}_t^a and $\hat{\mathbf{x}}_t^a$. Since $\hat{\mathbf{x}}_t^a$ is generated from \mathbf{x}_T^a , it follows the pipeline of existing diffusion model-based methods and tends to be a normal sample. On the contrary, \mathbf{x}_t^a is obtained by *directly* adding certain noises (i.e., t steps) to \mathbf{x}^a , and the anomalies will be preserved to some extent. In practice, considering that both $\hat{\mathbf{x}}_t^a$ and \mathbf{x}_t^a are with noises, and the noises may deviate from each other due to the denoising procedure, we propose to convert them to the noise-free version via Eq. (2), i.e., $\hat{\mathbf{x}}_{t \rightarrow 0}^a$ and $\mathbf{x}_{t \rightarrow 0}^a$. And the difference between $\hat{\mathbf{x}}_{t \rightarrow 0}^a$ and $\mathbf{x}_{t \rightarrow 0}^a$ is measured by the anomaly score, which is calculated according to Sec. 3.6. As shown in Fig. 2, if the difference is smaller than a threshold δ , then we continue the denoising steps (as shown in the green part of Fig. 2). Otherwise, if the difference appears (i.e., $> \delta$) at the t -th step (see the red part of Fig. 2), we can take \mathbf{x}_{t+n}^a as the starting point to take $t+n$ step of denoising, where n is a small number to preserve some redundancy. In this way, the details of normal regions are best preserved, and the anomalies can be reconstructed.

3.4 Spatial-Adaptive Feature Fusion

However, setting the redundant denoising step ($t+n$) for the whole image is sub-optimal. As analyzed in Sec. 3.2 and Eq. (5), we need only to set a larger step for the abnormal regions. Therefore, for normal regions, we can safely reduce the denoising step to t while keeping a large denoising step for the potential abnormal regions. As shown in Fig. 2, we have already performed the denoising steps from T , which can be reused in this procedure. Specifically, we can derive

a mask \mathbf{m} , which means the possibility for the pixels to be part of the anomalies. We pass the anomaly map in measuring difference between $\hat{\mathbf{x}}_{t \rightarrow 0}^a$ and $\mathbf{x}_{t \rightarrow 0}^a$, into the Sigmoid function as \mathbf{m} . Then, we can combine the two features with \mathbf{m} ,

$$\hat{\mathbf{x}}_t^f = \mathbf{m} \cdot \hat{\mathbf{x}}_t^a + (1 - \mathbf{m}) \cdot \mathbf{x}_t^a. \quad (6)$$

Note that the deviated noise problem still exists in Eq. (6), and we follow the strategy in Sec. 3.3 to add in the noise-free version and add the same noise ϵ following Eq. (1) for a consistent noise, *i.e.*,

$$\begin{aligned} \hat{\mathbf{x}}_t^f &= \sqrt{\bar{\alpha}_t} \hat{\mathbf{x}}_{t \rightarrow 0}^f + \sqrt{1 - \bar{\alpha}_t} \epsilon \\ &= \sqrt{\bar{\alpha}_t} (\mathbf{m} \cdot \hat{\mathbf{x}}_{t \rightarrow 0}^a + (1 - \mathbf{m}) \cdot \mathbf{x}_{t \rightarrow 0}^a) + \sqrt{1 - \bar{\alpha}_t} \epsilon. \end{aligned} \quad (7)$$

For a clear illustration, Fig. 2 is consistent with Eq. (6).

3.5 Anomaly-oriented Training Paradigm

With the modifications in Secs. 3.3 and 3.4, we have modulated the reconstruction process according to the properties of the anomaly detection task. However, as we can recall from Eq. (5), there still exists an incompatibility remaining unsolved. Particularly, better reconstruction quality implies a lower value of Eq. (5),

$$\hat{\mathbf{x}}^a - \mathbf{x} \approx \sqrt{1 - \bar{\alpha}_t} (\epsilon^a - \epsilon) + \sqrt{\bar{\alpha}_t} \mathbf{n} \rightarrow 0. \quad (8)$$

In Eq. (8), $\bar{\alpha}_t$, ϵ , and \mathbf{n} are manually set values or inherent concepts. The only value to be estimated by the model is ϵ^a . By rewriting Eq. (8), it should follow,

$$\epsilon^a \rightarrow \epsilon - \frac{\sqrt{\bar{\alpha}_t}}{\sqrt{1 - \bar{\alpha}_t}} \mathbf{n}. \quad (9)$$

Following the setting of diffusion models, $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is a random noise following standard Gaussian distribution. Once \mathbf{n} is non-zero (*i.e.*, anomalies exist), we can draw a conclusion that ϵ^a deviates from the standard Gaussian distribution.

For a typical diffusion model (ϵ_θ) trained with normal samples only, the noises in all steps follow the standard Gaussian distribution. In other words, it is beyond the scope of ϵ_θ to predict such an ϵ^a . Therefore, we propose introducing anomalies during training, which enables the diffusion models to break through the limitation of the standard Gaussian distribution and fit Eq. (9). On the basis of Eq. (9), the learning objective can be formulated by,

$$\begin{aligned} L_{ATP} &= \mathbb{E}_{(\mathbf{x}, \mathbf{x}^a) \sim p_{data}, \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), t} [\|(\epsilon - \frac{\sqrt{\bar{\alpha}_t}}{\sqrt{1 - \bar{\alpha}_t}} \mathbf{n}) - \epsilon^a\|_2] \\ &= \mathbb{E}_{(\mathbf{x}, \mathbf{x}^a) \sim p_{data}, \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), t} [\|(\epsilon - \frac{\sqrt{\bar{\alpha}_t}}{\sqrt{1 - \bar{\alpha}_t}} (\mathbf{x}^a - \mathbf{x})) - \epsilon_\theta(\mathbf{x}_t^a, t)\|_2]. \end{aligned} \quad (10)$$

One can see that Eq. (10) places even higher demands on the datasets than supervised anomaly detection since the corresponding normal sample is desired, which is difficult or even infeasible to prepare. To remedy the dilemma, we follow MemSeg [26] to synthesize abnormal samples with normal ones, which enables the training to proceed in an unsupervised manner.

It is obvious that the Eq. (10) is a more general form. For the normal region, the formula can degenerate to the original diffusion loss, while for the abnormal region, the formula forces the model to predict non-Gaussian noise.

3.6 Anomaly Scoring and Map Construction

Following AprilGAN [5], we regard reconstructed images as reference to compare with test images to construct anomaly maps. A pre-trained DINO [4] is utilized to extract the multi-layer features F_t of test images and F_r of reconstructed images, respectively. The anomaly map $M_l \in \mathbb{R}^{u \times v}$ of layer l are calculated based on the cosine similarity between layer l features $F_t^l \in \mathbb{R}^{c \times u \times v}$ and $F_r^l \in \mathbb{R}^{c \times u \times v}$, *i.e.*,

$$M_l^{(i,j)}(F_t^l, F_r^l) = \min(1 - \langle F_t^{l(i,j)}, F_r^l \rangle), \quad (11)$$

where (i, j) is the coordinate, $\langle x, \mathbf{y} \rangle$ calculates the cosine similarity between x and all elements of \mathbf{y} , and $\min(\cdot)$ returns the minimal value. Finally, the anomaly maps of different layers are added as the anomaly map $M \in \mathbb{R}^{u \times v}$, *i.e.*,

$$M = \sum_l M_l(F_t^l, F_r^l), \quad (12)$$

and the anomaly score is the average of the top K maximum values of M .

4 Experiments

In this section, we first introduce the experiment setup and details. Then, we compare our method with state-of-the-art (SOTA) methods to show the superiority of our method. In addition, To evaluate the effectiveness of our method, we conduct ablation studies on for proposed components.

4.1 Experiments Setup

Datasets. We conduct experiments on MVTec-AD [1] MPDD [11], VisA [33], and PCB-Bank datasets, to evaluate the effectiveness of our method.

MVTec-AD. MVTec-AD contains 10 objects and 5 texture classes of industrial anomalous samples with mask annotations. There are 3,629 images for training/validation and 1,725 images for testing.

MPDD. MPDD contains 6 classes of metal parts, comprising 888 normal samples for training and 458 samples either normal or anomalous for testing. Because of the variable spatial orientation, position, this dataset is challenging.

VisA. VisA is twice the size of MVTec, comprising 9,621 normal and 1,200 anomalous high-resolution images. This dataset exhibits objects of complex structures placed in sporadic locations and multiple objects in one image.

PCB-Bank. PCB-Bank is a printed circuit board dataset we integrated, including 7 different categories. There are 4214 normal samples for the training set and 2253 samples that are either normal or anomalous in the test set. The samples of the dataset have different clarity, resolution, and shooting angle.

Evaluation Metrics. Following prior works, we report on Area Under the Receiver Operating Curve (AUROC), Average Precision (AP) and F1-score-max (F1-max) in both anomaly detection and anomaly localization, where the prefix I- and prefix P- stand for anomaly detection and anomaly localization, respectively. Also, Per-Region-Overlap (PRO) is used in anomaly localization.

4.2 Implementation Details

We use the pre-trained latent diffusion model (LDM) [19] and fine-tune the UNet to adapt data. DINO [4] with ViT-B/8 architectures is utilized as a feature extraction model. To be consistent with the pre-trained VAE, images are resized to resolutions of 512×512 . We also report the results of 256×256 resolutions in Tab. 2. More details and multi-category settings parameter are included in the supplementary material.

4.3 Comparison with State-of-the-art Methods

Table 1: Comparison with SOTA methods on MVTec-AD dataset. I-AUROC and P-AUROC are displayed in each entry. The best results among all methods are shown in bold. The best results among reconstruction-based methods are underlined.

Category	Embedding-based methods			Reconstruction-based methods					
	PatchCore [20]	RD4AD [7]	SimpleNet [16]	DRAEM [28]	OCR-GAN [13]	Lu <i>et al.</i> [17]	DiffAD [29]	DDAD [18]	Ours
Carpet	98.7/ 99.0	98.9/98.8	99.7 /98.2	97.0/95.5	<u>99.4</u> /-	-/97.7	98.3/98.1	99.3/ <u>98.7</u>	99.0/98.5
Grid	98.2/98.7	100 /97.0	99.7/98.8	99.9/ 99.7	99.6/-	-/95.6	100 / 99.7	<u>100</u> /99.4	100 /99.6
Leather	100 /99.3	100 /98.6	100 /99.2	100 /98.5	97.1/-	-/97.5	100 /99.1	100 /99.4	100 / 99.8
Tile	98.7/95.6	99.3/98.9	99.8/97.0	99.6/99.2	95.5/-	-/98.9	100 / 99.4	100 /98.2	100 /98.7
Wood	99.2/95.0	99.2/ 99.3	100 /94.5	99.1/96.4	95.7/-	-/99.1	100 /96.7	100 /95.0	99.4/98.4
Bottle	100 /98.6	100 /99.0	100 /98.0	99.2/ 99.1	99.6/-	-/97.3	100 /98.8	100 /98.7	100 /98.9
Cable	99.5/98.4	95.0/99.4	99.9 /97.6	91.8/94.7	99.1/-	-/99.5	94.6/96.8	99.4/98.1	99.9 /98.1
Capsule	98.1/98.8	96.3/97.3	97.7/ 98.9	98.5/94.3	96.2/-	-/96.8	97.5/98.2	99.4/95.7	99.5 / <u>98.5</u>
Hazelnut	100 /98.7	99.9/98.2	100 /97.9	100 / 99.7	98.5/-	-/92.5	100 /99.4	100 /98.4	100 /99.5
Metal nut	100 /98.4	100 / 99.6	100 /98.8	98.7/ <u>99.5</u>	99.5/-	-/99.0	100 /99.4	100 /99.0	100 /98.8
Pill	96.6/97.4	99.6 /95.7	99.0/ 98.6	98.9/97.6	98.3/-	-/92.1	97.7/97.7	100 /99.1	98.1/ <u>97.9</u>
Screw	98.1/ 99.4	97.0/99.1	98.2/99.3	93.9/97.6	100 /-	-/98.6	97.2/99.0	99.0/ <u>99.3</u>	96.9/99.1
Toothbrush	100 /98.7	99.5/93.0	99.7/98.5	100 /98.1	98.7/-	-/93.1	100 /99.2	100 /98.7	100 / 99.4
Transistor	100 /96.3	96.7/95.4	100 / 97.6	93.1/90.9	<u>98.3</u> /-	-/94.5	96.1/93.7	100 /95.3	98.3/ <u>96.2</u>
Zipper	99.4/98.8	98.5/98.2	99.9/98.9	100 /98.8	99.0/-	-/97.6	100 / 99.0	100 /98.2	98.5/97.9
Average	99.1/98.1	98.5/97.8	99.6/98.1	98.0/97.3	98.3/-	-/96.7	98.7/98.3	99.8 /98.1	99.3/ 98.6

Comparisons with state-of-the-art (SOTA) methods on the MVTec-AD dataset are shown in Tab. 1. The methods include embedding-based methods (PatchCore [20], RD4AD [7] and SimpleNet [16]), and reconstruction-based methods (DRAEM [28], OCR-GAN [13], Lu *et al.* [17], DiffAD [29] and DDAD [18]). Lu *et al.*, DiffAD, and DDAD are advanced diffusion-based methods. For image-level anomaly detection tasks, our method achieves the highest I-AUROC on 9 out of 15 classes. Although the I-AUROC of our method is slightly lower than DDAD’s, our method surpasses DDAD on I-AP and I-F1-max (99.7/98.4 VS 99.5/97.9) in Tab. 2. For anomaly localization tasks, our method outperforms the SOTA among all types of methods, and exceeds reconstruction based SOTA (DDAD) by 11.9 \uparrow /9.6 \uparrow /3.0 \uparrow in P-AP/P-F1-max/PRO.

We also conduct experiments on the MPDD dataset in Tab. 2. Our method outperforms the SOTA among all methods, and surpasses the reconstruction-based SOTA by 1.8 \uparrow /10.5 \uparrow /5.7 \uparrow /5.7 \uparrow on P-AUROC/P-AP/P-F1-max/PRO. The SOTA also achieved by GLAD on VisA and PCB-Bank datasets. On the average of the four datasets, GLAD outperforms existing methods on all metrics.

Table 2: Quantitative results on MVTec-AD, MPDD, VisA and PCB-Bank datasets. Metrics are I-AUROC/I-AP/I-F1-max at first row (for detection) and P-AUROC/P-AP/P-F1-max/PRO at second row (for localization).

Dataset	MVTec-AD	MPDD	VisA	PCB-Bank	Avg
PatchCore [20]	99.1/99.6/98.1	91.3/95.1/91.3	91.0/92.7/88.7	94.2/95.6/90.3	93.9/95.8/92.1
	98.1/55.9/57.6/93.4	98.5/38.4/40.7/92.9	98.1/38.5/40.5/88.3	99.1/46.0/48.5/90.8	98.5/44.7/46.9/91.4
RD4AD [7]	98.7/99.5/98.0	95.3/96.8/93.0	96.9/97.2/93.8	96.0/96.2/92.6	96.8/97.5/94.4
	97.9/59.0/61.2/94.1	98.7/44.5/46.1/95.2	98.3/44.6/47.2/93.0	99.1/46.3/48.0/94.0	98.5/48.9/50.8/94.2
SimpleNet [16]	99.6/99.6/98.9	96.6/97.7/96.0	96.2/96.9/92.6	97.4/98.1/94.6	97.5/98.1/95.5
	98.1/49.8/52.8/91.9	97.4/35.6/37.5/90.4	98.5/33.2/37.1/92.3	99.0/43.3/45.2/94.4	98.3/40.5/43.2/92.3
DRAEM [28]	98.0/99.0/96.9	94.3/95.8/93.0	92.4/93.4/87.9	91.5/91.8/88.2	94.1/95.0/91.5
	97.3/68.4/66.7/91.3	90.7/28.3/29.8/78.0	92.0/28.8/36.1/78.7	96.4/32.2/38.0/80.9	94.1/39.4/42.7/82.2
OCR-GAN [13]	98.3/98.1/95.0	96.2/96.6/97.7	97.9/98.7/96.4	91.3/91.6/88.0	95.9/96.3/94.3
	-/-/-	-/-/-	-/-/-	-/-/-	-/-/-
DDAD [18]	99.8/99.5/97.9	97.2/95.5/95.1	98.9/98.6/96.2	97.4/96.1/95.2	98.3/97.4/96.1
	98.1/59.0/59.4/92.3	96.9/34.8/43.5/91.6	97.6/27.9/34.6/92.7	96.5/28.1/33.6/91.1	97.3/37.5/42.8/91.9
Ours-256	99.0/99.7/98.2	97.3/98.4/95.4	99.3/99.6/97.6	98.1/98.4/96.7	98.4/98.8/97.0
	98.7/63.8/63.7/95.2	98.5/41.5/43.9/94.2	98.3/35.8/42.4/94.1	98.8/42.0/47.5/93.8	98.6/44.5/48.6/94.3
Ours-512	99.3/99.7/98.4	97.5/98.5/96.0	99.5/99.4/98.3	98.7/99.0/97.3	98.8/99.2/97.5
	98.6/70.9/69.0/95.3	98.7/45.3/49.2/96.3	98.6/39.1/45.4/94.3	99.3/48.9/52.2/95.1	98.8/51.1/54.0/95.3

Reconstructions and qualitative results on the MVTec-AD and MPDD datasets are displayed in Fig. 3. More quantitative results are presented in supplementary materials. Other methods usually fail to reconstruct large-scale anomalies into normal regions and produce inaccurate locations. On the contrary, our method can produce satisfactory reconstruction and accurately location.

More details and results of multi-class settings are in supplementary material.

4.4 Ablation Study

In this section, ablation studies are conducted to verify the effectiveness of our proposed components: Adaptive Denoising Steps (ADS), Spatial-Adaptive Feature Fusion (SAFF) and Anomaly-oriented Training Paradigm (ATP). A discussion of hyperparameters is presented in the supplementary materials.

Adaptive Denoising Steps With the ADS, LDM chooses the proper steps and denoises from the corresponding noisy samples, which ensures anomaly-free reconstruction and preserves as much normal information as possible. There is the obvious improvement in Tab. 3. In Tab. 4, we compare ADS with other cases which use different fixed steps. The results prove the superiority of ADS.

Table 3: Performance of each component on MVTec-AD dataset.

Method	I-AUROC	P-AUROC
Baseline (LDM)	98.3	98.0
Baseline + ADS	99.0	98.5
Baseline + ATP	98.7	98.5
Baseline + ADS + ATP w/o SAFF	99.2	98.3
Baseline + ADS + ATP with SAFF (Ours)	99.3	98.6

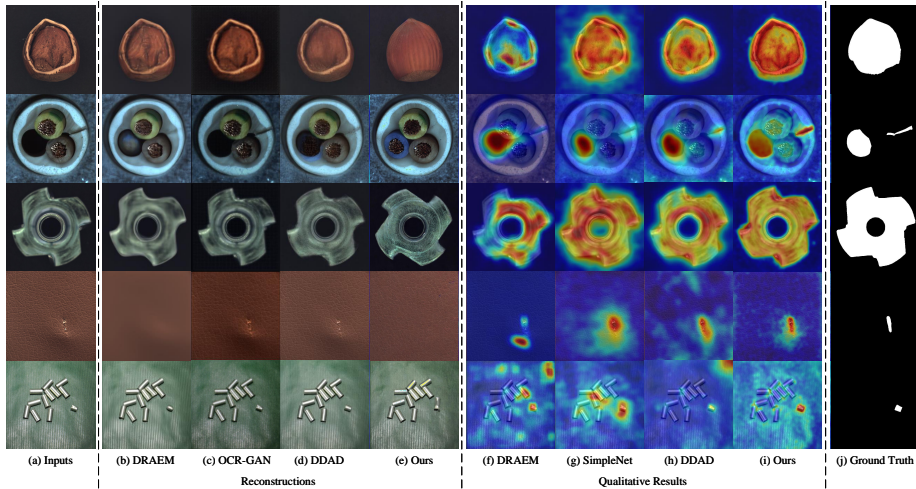


Fig. 3: Reconstructions and qualitative comparisons with other methods. The first four rows display examples of the MVTec-AD dataset, and the last row is for the MPDD dataset. OCR-GAN only produces anomaly scores, and there is no anomaly map. SimpleNet is the embedding-based method.

Table 4: Comparison of adaptive steps and different fixed steps on MVTec-AD dataset.

Denoising steps	fixed steps			Adaptive steps
	350 step	550 step	750 step	
I-AUROC/P-AUROC	98.8/97.6	98.8/98.1	98.7/98.5	99.3/98.6

We display different types of anomaly and proper steps chosen by ADS in Fig. 4. For anomalies that are not obvious, such as examples (a), (c), and (e), ADS tends to select small steps that are enough for reconstruction. However, for large-scale anomalies, such as examples (b), (d), and (f), ADS select larger step noise to ensure the reconstruction of anomaly-free images.

Besides, we observed that ATP can further reduce the denoising steps. Fig. 4 displays the reconstructions of ADS and the combination of ADS and ATP. ATP can help the model better remove anomalies during inference. Thus, the clear image for comparison will contain less anomaly at each step, and the difference is detected in smaller steps. This retains more normal details to produce a more accurate reconstruction. The details are marked with red circles in Fig. 4. The reconstruction of ADS with ATP is more similar with inputs at normal areas. Performance can be further improved as shown in Tab. 3.

Spatial-Adaptive Feature Fusion In SAFF, we fuse the features from the predicted sample and the test sample with a mask m , which is generated with

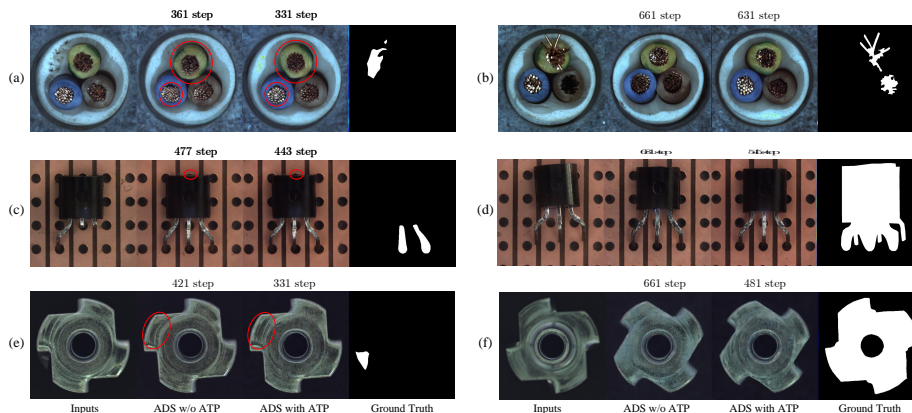


Fig. 4: Reconstructions of different types of anomaly and proper steps. Examples (a), (c), and (e) contain small-scale anomalies, and (b), (d), and (f) are large-scale anomalies. The numbers above the reconstructed images represent the proper steps. Differences in details of normal areas are marked in red circles.

the anomaly map at the proper denoising step. As shown in Tab. 3, SAFF can further remove residual anomaly information and improve performance.

Anomaly-oriented Training Paradigm Because of adding synthesis anomaly samples in training, the data distribution variance between the training and test data is narrowed by our ATP. Thus, LDM can map abnormal regions as normal regions as well. Performance of baseline and ATP are shown in Tab. 3. There are visual comparisons in Fig. 5. For both baseline and ATP, samples are added same noise and reconstructed without ADS. In the column (b) and (e), the results of the baseline still contain anomaly. This suggests that the reconstruction ability of the baseline is limited. After training with ATP, as shown in column (c) and (f), the model can map abnormal regions into normal regions well.

5 Social Impact, Limitations, and Future Work

This work studies the problem of unsupervised anomaly detection and achieves SOTA performance on four datasets. Without the need for real-world abnormal samples, this work has the potential to be efficiently utilized in real-world scenarios. Despite the appealing performance, our method introduces evaluation comparison in inference, which causes extra time costs. Besides, to determine the denoising step, we conduct the comparison step-by-step, and the denoising steps are not actually reduced. In future work, we plan to design a lightweight evaluation comparison and predict the denoising step with a limited number of denoising steps, which has the potential to improve the efficiency of our method.

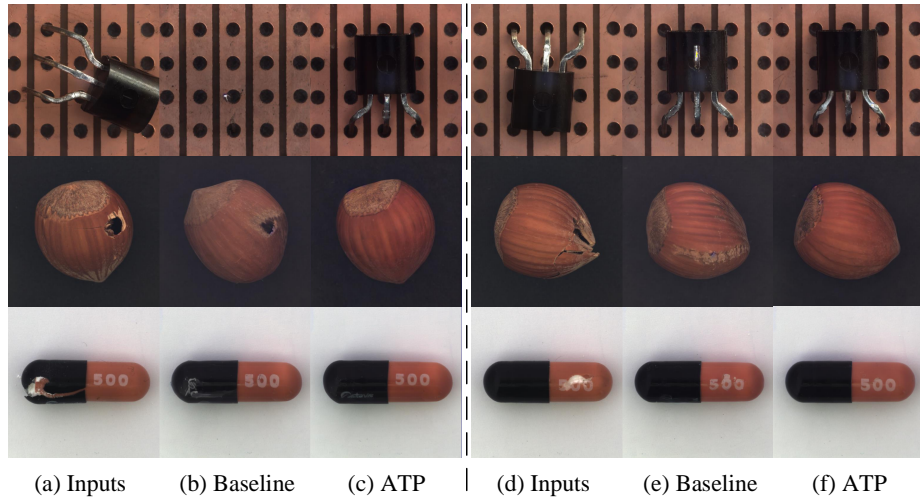


Fig. 5: Qualitative comparisons between baseline and proposed ATP on MVTec-AD. The same denoising steps are used for the two methods.

6 Conclusion

In this paper, a global and local adaptive diffusion model is presented to improve reconstruction results. As a global configuration, an adaptive denoising step is set for each sample to adapt to different anomalies. Moreover, considering that reconstructing the abnormal regions is different from the normal areas, an adaptive feature fusion scheme is proposed to remove residual anomalies, and an anomaly-oriented training paradigm is proposed to promote the reconstruction ability of DM, which achieves local adaptive reconstructions. We conduct extensive experiments on MVTec-AD, MPDD, VisA, and PCB-Bank datasets. Quantitative and qualitative results evaluate the superiority of our approach.

Acknowledgement

This work was supported in part by the National Key Research and Development Program of China under Grant No. 2023YFA1008500.

References

1. Bergmann, P., Fauser, M., Sattlegger, D., Steger, C.: Mvtec ad—a comprehensive real-world dataset for unsupervised anomaly detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9592–9600 (2019)

2. Bergmann, P., Fauser, M., Sattlegger, D., Steger, C.: Uninformed students: Student-teacher anomaly detection with discriminative latent embeddings. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4183–4192 (2020)
3. Bergmann, P., Löwe, S., Fauser, M., Sattlegger, D., Steger, C.: Improving unsupervised defect segmentation by applying structural similarity to autoencoders. arXiv preprint arXiv:1807.02011 (2018)
4. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9650–9660 (2021)
5. Chen, X., Han, Y., Zhang, J.: A zero-/few-shot anomaly classification and segmentation method for cvpr 2023 vand workshop challenge tracks 1&2: 1st place on zero-shot ad and 4th place on few-shot ad. arXiv preprint arXiv:2305.17382 (2023)
6. Defard, T., Setkov, A., Loesch, A., Audigier, R.: Padim: a patch distribution modeling framework for anomaly detection and localization. In: International Conference on Pattern Recognition. pp. 475–489. Springer (2021)
7. Deng, H., Li, X.: Anomaly detection via reverse distillation from one-class embedding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9737–9746 (2022)
8. Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A.H., Chechik, G., Cohen-Or, D.: An image is worth one word: Personalizing text-to-image generation using textual inversion. arXiv preprint arXiv:2208.01618 (2022)
9. He, H., Zhang, J., Chen, H., Chen, X., Li, Z., Chen, X., Wang, Y., Wang, C., Xie, L.: Diad: A diffusion-based framework for multi-class anomaly detection. arXiv preprint arXiv:2312.06607 (2023)
10. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. *Advances in neural information processing systems* **33**, 6840–6851 (2020)
11. Jezek, S., Jonak, M., Burget, R., Dvorak, P., Skotak, M.: Deep learning-based defect detection of metal parts: evaluating current methods in complex conditions. In: 2021 13th International congress on ultra modern telecommunications and control systems and workshops (ICUMT). pp. 66–71. IEEE (2021)
12. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
13. Liang, Y., Zhang, J., Zhao, S., Wu, R., Liu, Y., Pan, S.: Omni-frequency channel-selection representations for unsupervised anomaly detection. *IEEE Transactions on Image Processing* (2023)
14. Liu, W., Li, R., Zheng, M., Karanam, S., Wu, Z., Bhanu, B., Radke, R.J., Camps, O.: Towards visually explaining variational autoencoders. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8642–8651 (2020)
15. Liu, X., Wei, Y., Liu, M., Lin, X., Ren, P., Xie, X., Zuo, W.: Smartcontrol: Enhancing controlnet for handling rough visual conditions. arXiv preprint arXiv:2404.06451 (2024)
16. Liu, Z., Zhou, Y., Xu, Y., Wang, Z.: Simplenet: A simple network for image anomaly detection and localization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20402–20411 (2023)
17. Lu, F., Yao, X., Fu, C.W., Jia, J.: Removing anomalies as noises for industrial defect localization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 16166–16175 (2023)
18. Mousakhan, A., Brox, T., Tayyub, J.: Anomaly detection with conditioned denoising diffusion models. arXiv preprint arXiv:2305.15956 (2023)

19. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
20. Roth, K., Pemula, L., Zepeda, J., Schölkopf, B., Brox, T., Gehler, P.: Towards total recall in industrial anomaly detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14318–14328 (2022)
21. Rudolph, M., Wehrbein, T., Rosenhahn, B., Wandt, B.: Asymmetric student-teacher networks for industrial anomaly detection. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 2592–2602 (2023)
22. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22500–22510 (2023)
23. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: International conference on machine learning. pp. 2256–2265. PMLR (2015)
24. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: International Conference on Learning Representations (2020)
25. Wei, Y., Zhang, Y., Ji, Z., Bai, J., Zhang, L., Zuo, W.: Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15943–15953 (2023)
26. Yang, M., Wu, P., Feng, H.: Memseg: A semi-supervised method for image surface defect detection using differences and commonalities. *Engineering Applications of Artificial Intelligence* **119**, 105835 (2023)
27. Yin, H., Jiao, G., Wu, Q., Karlsson, B.F., Huang, B., Lin, C.Y.: Lafite: Latent diffusion model with feature editing for unsupervised multi-class anomaly detection. arXiv preprint arXiv:2307.08059 (2023)
28. Zavrtnik, V., Kristan, M., Skočaj, D.: Draem-a discriminatively trained reconstruction embedding for surface anomaly detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8330–8339 (2021)
29. Zhang, X., Li, N., Li, J., Dai, T., Jiang, Y., Xia, S.T.: Unsupervised surface anomaly detection with diffusion probabilistic model. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6782–6791 (2023)
30. Zhang, Y., Wei, Y., Jiang, D., ZHANG, X., Zuo, W., Tian, Q.: Controlvideo: Training-free controllable text-to-video generation. In: The Twelfth International Conference on Learning Representations (2023)
31. Zhang, Y., Wei, Y., Lin, X., Hui, Z., Ren, P., Xie, X., Ji, X., Zuo, W.: Videoelevator: Elevating video generation quality with versatile text-to-image diffusion models. arXiv preprint arXiv:2403.05438 (2024)
32. Zhao, Y., Deng, B., Shen, C., Liu, Y., Lu, H., Hua, X.S.: Spatio-temporal autoencoder for video anomaly detection. In: Proceedings of the 25th ACM international conference on Multimedia. pp. 1933–1941 (2017)
33. Zou, Y., Jeong, J., Pemula, L., Zhang, D., Dabeer, O.: Spot-the-difference self-supervised pre-training for anomaly detection and segmentation. In: European Conference on Computer Vision. pp. 392–408. Springer (2022)