# Removing Anomalies as Noises for Industrial Defect Localization

Fanbin Lu[1] Xufeng Yao[1] Chi-Wing Fu[1] Jiaya Jia [1,2]
[1]The Chinese University of Hong Kong [2]SmartMore
{fblu21, xfyao, cwfu, leojia}@cse.cuhk.edu.hk

## Abstract

*Unsupervised anomaly detection aims to train models with only anomaly-free images to detect and localize unseen anomalies. Previous reconstruction-based methods have been limited by inaccurate reconstruction results. This work presents a denoising model to detect and localize the anomalies with a generative diffusion model. In particular, we introduce random noise to overwhelm the anomalous pixels and obtain pixel-wise precise anomaly scores from the intermediate denoising process. We find that the KL divergence of the diffusion model serves as a better anomaly score compared with the traditional RGB space score. Furthermore, we reconstruct the features from a pre-trained deep feature extractor as our feature level score to improve localization performance. Moreover, we propose a gradient denoising process to smoothly transform an anomalous image into a normal one. Our denoising model outperforms the state-of-the-art reconstruction-based anomaly detection methods for precise anomaly localization and high-quality normal image reconstruction on the MVTec-AD benchmark.*

## 1. Introduction

Anomaly detection is a critical computer vision task that has great application values in industry and medicine. Despite its importance, collecting and annotating anomalous data can be prohibitively expensive. Unsupervised anomaly detection recently garnered significant attention. Different from few-shot segmentation [12, 33, 32], it aims to learn normal data distribution without access to anomalous samples and ground-truth annotations in training. At inference, anomalies are detected and localized based on their deviation from the learned distribution of normal data.

Classical reconstruction-based unsupervised anomaly detection methods [1, 2, 5, 20] assume the autoencoder model trained with only normal data fail to reconstruct anomalous regions. However, this approach is not without limitations, as some anomalies can still be reconstructed, leading to the inferior performance of these classical methods. DRAEM [40] proposes to generate pseudo anomalies to train an autoencoder to reconstruct the anomalous data to be anomaly-free. However, it performs poorly when the real anomalies differ significantly from the pseudo ones. Denoising autoencoders [17] are used for medical anomaly detection. The anomaly score is measured naively by the difference between the input and reconstructed images in pixel space. The reconstruction from noisy images is challenging and introduces great noise to the results, making it unsuitable for complex industrial anomaly detection and localization. Recently, methods [18, 22, 36, 37] propose transformer structures for the reconstruction model to prevent the autoencoder from collapsing into an identity func-

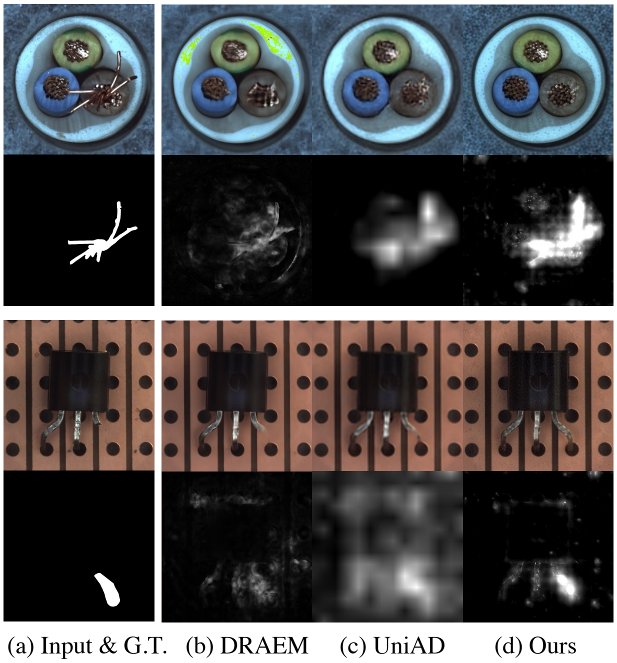


Figure 1: Comparing reconstructed normal images (rows 1 & 3) and anomaly detection results (rows 2 & 4) produced by different methods. Our method can produce high-quality reconstructions without obvious artifacts (DRAEM [40]) and blurring (UniAD [36]) and locate anomalies more precisely.

tion. The models produce blurred results with relatively poor anomaly localization performance compared with the state-of-the-art.

In this work, we propose a new reconstruction-based approach for anomaly detection, achieving precise anomaly localization and top reconstruction quality; see Fig. 1. Our key idea is to formulate the anomaly detection task as a noise or anomaly removal problem. First, we introduce random noises into the input image and train an autoencoder as a denoising model. The anomalous pixels are considered as noises and will not be excluded from the reconstruction. The previous reconstruction-based methods directly reconstruct the input with noisy images, leading to large reconstruction errors and suboptimal anomaly detection performance. We leverage a diffusion model [15] for denoising and reconstruction. We inspect the intermediate stages of the diffusion model and measure the reconstruction error of each step for accurate anomaly localization. Moreover, we require the model to reconstruct the input features and detect anomalies in both pixel and feature space.

We further propose a gradient denoising process for reconstructing normal images from anomalous ones and provide an interpretable explanation of the anomaly detection results. Our process smoothly transforms an anomalous image into a normal image while preserving the structural appearance and high-frequency details of the normal regions. It is achieved by consistently denoising the gradients from a pre-trained deep feature extractor. Our approach is shown to outperform existing methods in terms of reconstruction quality and anomaly detection accuracy.

## 2. Related Works

**Anomaly Detection** Various methods have been developed to tackle anomaly detection and localization. Support vector data description (SVDD) [31, 26] is proposed for anomaly detection. Teacher-Student [4] proposes to distill the knowledge from a pre-trained teacher network to a student network on the anomaly-free data. The difference in the outputs of teacher and student networks is used as an anomaly score. DRAEM [40] proposes to add artificial defects to the normal images to generate pseudo anomaly samples and labels to train a segmentation network for anomaly segmentation. CutPaste [19] proposed a self-training strategy with a generative one-class classifier.

Reconstruction-based approaches [5, 7] are a widely used branch of anomaly detection. They assume that only the normal image can be well reconstructed. Anomalies can be detected by measuring the difference between original and reconstructed images. Autoencoders [5, 7], variational autoencoders (VAE) [34], and Adversarial generative networks (GAN) [1] are often used to reconstruct an anomalous image to a normal one. However, a limitation of these methods is that anomalies can sometimes be reconstructed, leading to degraded anomaly detection performance.

Embedding-based methods [6, 8, 25] employ neural networks to extract meaningful features for anomaly detection and localization. Spade [6] first introduced a method for detecting anomalies using ImageNet pre-trained deep networks. This method uses K-NN search to match anomaly features with the K nearest normal features. PaDiM [8] build a multivariate Gaussian distribution and use Mahalanobis distance as the anomaly score. PatchCore [25] proposes a memory bank to save the coreset of the normal features, which improves the time and memory complexity. Recently, UniAD [36] proposed a transformer network for reconstructing features with masked self-attention to avoid the model collapsing into an identity function. This allows a single model to detect anomalies in all categories.

Flow-based methods [11, 16, 13, 38] recently boosted the performance of anomaly detection. Normalized flow models are generative models that learn to map two distributions and estimate the probability density reversibly. CFLOW-AD [13] proposes to use conditional normalized flow with positional embedding on the multi-scale features for anomaly detection. FastFlow [38] proposes to employ a 2D flow model that combines local and global features to estimate the probability density. These methods demonstrate the efficacy of generative models in addressing anomaly detection, which has inspired our work with the diffusion model.

**Diffusion Models** Diffusion models [27, 15] are a powerful generative model that achieves state-of-the-art performance in image generation tasks. Recent methods [15, 23, 10] are proposed to generate a realistic image by gradually denoising random Gaussian noises. The likelihood training makes the diffusion model capable of learning data density. DDIM [28] speeds up the diffusion sampling with a non-Markov reverse sampling. The score-based model [29] is another denoising generative model with a similar diffusion process.

AnoDDPM [35] has introduced diffusion models for medical image anomaly segmentation but only used the diffusion model as a high-quality reconstruction model. Relying on reconstruction error in RGB space for anomaly score leads to noisy predictions and limited performance in many industrial anomaly detection applications.

## 3. Methods

### 3.1. Preliminary

Diffusion models are powerful generative models that can approximate data distribution and create realistic images. Given a data distribution $p(x)$, denoising diffusion probabilistic models (DDPM) [15] learn the distribution with a Markov Chain denoising process. During training, it gradually adds random Gaussian noises to a real image $\boldsymbol{x}_0$,

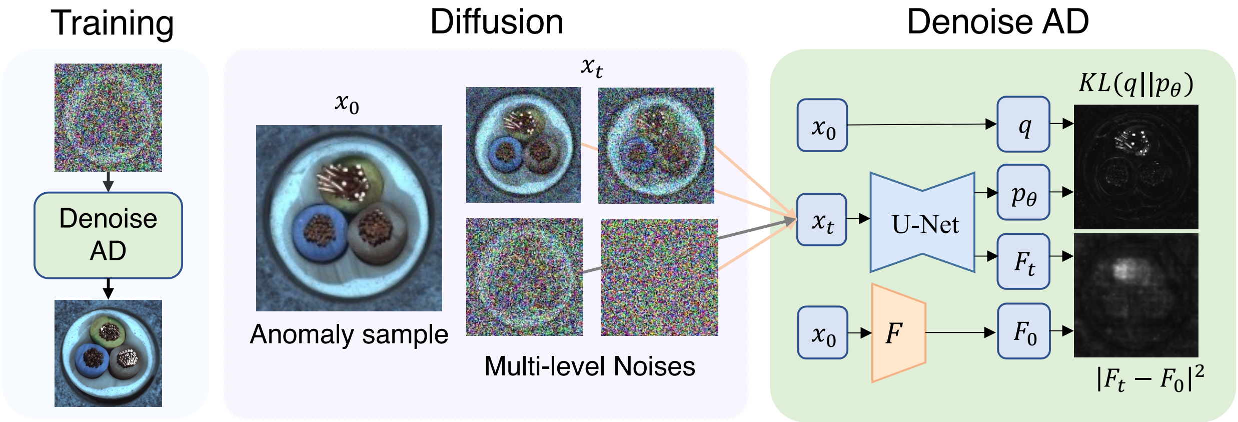


Figure 2: Our denoising diffusion model is trained with only anomaly-free images. During inference, noises of different scales are added to the anomaly sample. With large enough noises, the anomalous pixels become indistinguishable from the normal pixels and easier for reconstruction. We take the KL-divergence between the posterior distribution $q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)$ and estimated distribution $p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$ as the pixel-level anomaly score. The MSE error of feature reconstruction is used as a feature-level score. We take the average of results from different noise scales as the outputs.

which leads to a series of noised images $(\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_T)$. The variances of Gaussian noises introduced are denoted as $\{\beta_t\}_{t=1,2,\cdots,T}$. Since the data distribution and noises added are both Gaussian, the closed form of a noised image $\boldsymbol{x}_t$ is:

$$q(\boldsymbol{x}_t|\boldsymbol{x}_0) = \mathcal{N}(\boldsymbol{x}_t; \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0, (1-\bar{\alpha}_t)\boldsymbol{I}), \tag{1}$$

where $\bar{\alpha}_t = \prod_{s=1}^{t}(1-\beta_s)$. The diffusion models are then represented with $p_\theta(\boldsymbol{x}_0) = \int p(\boldsymbol{x}_{0:T})\,\mathrm{d}\boldsymbol{x}_{1:T}$. During the image generation process, the model first samples from uniform Gaussian distribution $p_T(\boldsymbol{x}) \sim \mathcal{N}(\boldsymbol{x}_T; \boldsymbol{0}, \boldsymbol{I})$ and gradually denoises the image by sampling from the estimated distribution $p_\theta(\boldsymbol{x}_t)$:

$$p_\theta(\boldsymbol{x}_{0:T}) = p(\boldsymbol{x}_T)\prod_{t=1}^{T} p(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t), \tag{2}$$

$$p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t) = \mathcal{N}(\boldsymbol{x}_{t-1}; \boldsymbol{\mu}_\theta(\boldsymbol{x}_t, t), \boldsymbol{\Sigma}_\theta(\boldsymbol{x}_t, t)). \tag{3}$$

The training of diffusion models can be treated as an autoencoder. As proposed in DDPM [15], the diffusion models are trained with an MSE loss to predict the scale of noises $\epsilon$.

$$L_{mse} = \mathbb{E}_{t,\boldsymbol{x}_0,\boldsymbol{\epsilon}}\left[(\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t))^2\right] \tag{4}$$

An additional training loss based on the variational bound is used to automatically learn the variance of noises by the diffusion model itself, as proposed by [23]:

$$L_{vlb} = L_0 + L_1 + \cdots + L_{T-1} + L_T, \tag{5}$$

$$L_0 = -logp_\theta(\boldsymbol{x}_0|\boldsymbol{x}_1), \tag{6}$$

$$L_{t-1} = D_{KL}(q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)||p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)), \tag{7}$$

$$L_T = D_{KL}(q(\boldsymbol{x}_T|\boldsymbol{x}_0)||p(\boldsymbol{x}_T)). \tag{8}$$

### 3.2. Denoising Model for Anomaly detection

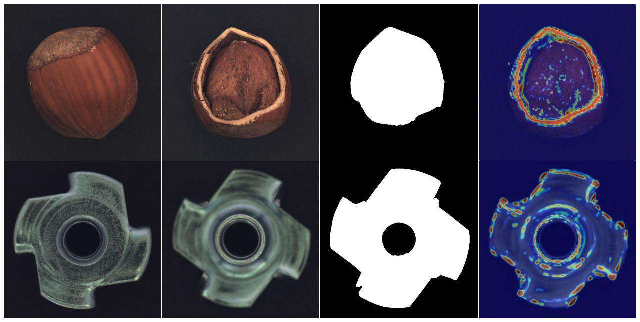
Previous reconstruction methods based on AutoEncoder [2, 5] suffer from the successful reconstruction of anomalies because the AutoEncoder easily degrades to an identical mapping during training. However, reconstruction with noisy images prevents the issue. As illustrated in Fig 2, gradually adding noise to an anomalous image causes the anomalous regions to vanish for large noise levels, making them indistinguishable from the pixels of normal samples. Nevertheless, direct reconstruction from noisy to noise-free images can result in significant reconstruction errors. In this study, we utilize a generative diffusion model DDPM [15] to gradually denoise and reconstruct the image. The diffusion model is trained on anomaly-free data using the training procedure of DDPM.

**Pixel-level score.** For anomaly detection, we begin by corrupting an image $\boldsymbol{x}_0$ with random Gaussian noises to obtain $\boldsymbol{x}_t$. Previous reconstruction-based methods employ the difference between the reconstructed image and the original input in RGB space as the anomaly score. However, this approach entails a difficult estimation of $p(\boldsymbol{x}_0|\boldsymbol{x}_t)$ and introduces significant noise to the results. To address this limitation, we employ the KL divergence of the posterior distribution $q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)$ and the estimated distribution $p\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$ as the anomaly score,

$$\boldsymbol{s}_t = KL(q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)||p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)). \tag{9}$$

We show in Fig. 6 that the KL divergence correctly measures the likelihood of input pixels with much less noise.

**Feature-level score.** We observe that the results from the diffusion model are usually sharp in boundary but not robust



(a) Normal (b) Anomalous (c) Label (d) Prediction

Figure 3: Ambiguity of label. The diffusion model focuses on the anomalous pixels that need to be altered for successful reconstruction. It requires semantic information to correctly segment large-area hazelnut cracking and metal nut flipping.

in anomaly recall. As demonstrated in Fig. 3, the anomaly regions with a similar color to the normal pixels are assigned with a high likelihood by the denoising model. The diffusion model prioritizes the anomalous pixels that need to be altered for successful reconstruction, which requires semantic information to address the issue.

To enhance the accuracy of anomaly detection, we propose a joint distribution approach that considers both the pixel space and feature space, represented by $P(\boldsymbol{x}, \boldsymbol{f})$. We employ a pre-trained feature extractor to extract the deep features of the input image. The diffusion model is trained to concurrently reconstruct the pixels and semantic features of the noise-free image with the corrupted image. We adopt the Mean Squared Error (MSE) loss function as the training loss and anomaly score, which is defined as follows:

$$s_t^f = L_{mse}^f = \frac{1}{C \times H \times W} \sum |f(\boldsymbol{x}_0) - f(\boldsymbol{x}_t)|^2, \quad (10)$$

where $f$ is a pre-trained feature extractor to extract features with shape $\mathbb{R}^{C \times H \times W}$, $\boldsymbol{x}_0$ and $\boldsymbol{x}_t$ represent a noise-free image and the corresponding corrupted image with random noises, respectively. The final anomaly score is the weighted sum of the pixel-level and feature-level results.

**Multi-scale noises.** We have observed that different anomalies exhibit varying sensitivities to different noise scales. While some anomalies can be detected easily, others require sufficiently large noise to overwhelm the anomalous pixels. We measure the anomaly score for various noise scales and average the results. Since the KL-divergence score varies significantly with the timestep $t$, we normalize it before averaging. The final anomaly score is obtained as follows:

$$A = \sum_{i=1,2,\ldots,n} \alpha \hat{s}_{t_i} + (1-\alpha) s_{t_i}^f, \quad (11)$$

**Algorithm 1:** Gradient Denoising Reconstruction.

```
Input: Image $\boldsymbol{x}_0$, Gaussian $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
Output: $\boldsymbol{x}_N$
for $t = 1, \cdots, N$ do
    $\boldsymbol{f}_t = F(\boldsymbol{x}_t)$
    $\boldsymbol{g} = \nabla_{\boldsymbol{x}_t} (\boldsymbol{f}_t - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{f}_t - \boldsymbol{\mu})$
    $\boldsymbol{x}_t = \sqrt{1 - \hat{\beta}_t} \boldsymbol{x}_{t-1} + \sqrt{\hat{\beta}_t} \boldsymbol{g}$ ;
    if $t \% N_d = 0$ then
        $\boldsymbol{x}_t \sim \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{\theta}}(\boldsymbol{x}_t), \boldsymbol{\sigma}_{\theta}(\boldsymbol{x}_t))$
    end
end
```

where $\hat{s}_{t_i}$ is the normalized score by mean and standard deviation, $T = \{t_1, t_2, \cdots, t_n\}$ are the selected timesteps of the forward-process of the diffusion model. We analyze the effects of ensembling factor $\alpha$ in Sec. 4.4.

**Unified model.** It has been proved that the diffusion model's capacity of the diffusion network is large enough for modeling any complex distributions [15, 10]. Like UniAD [36], we conduct experiments to learn distributions of multiple categories with a single diffusion model. Table 3 illustrates that the performance of our unified model outperforms the other methods by a large margin under the single unified model setting. The results confirm the effectiveness of utilizing the diffusion model for anomaly localization.

### 3.3. Gradient Denosing for Reconstruction

An image's anomalous regions can be viewed as a special type of noise that can be removed using the diffusion model. We propose a gradient denoising process to remove the anomalies with simple adjustments to the reverse diffusion process of DDPM [15]. An anomalous image can be smoothly transformed into a normal one, providing an interpretable explanation of the anomaly detection results.

We first introduce a gradient descending optimization process for anomaly reconstruction. We take the multivariate Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ approximated by PaDiM [8] on the deep features of anomaly-free data. For reconstruction, we extract embedding $f(\boldsymbol{x}_0)$ with the feature extractor of PaDiM and use the Mahalanobis distance to optimize the image with gradient descending:

$$L = (f(\boldsymbol{x}_0) - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (f(\boldsymbol{x}_0) - \boldsymbol{\mu}), \quad (12)$$

$$\boldsymbol{x}_{t+1} = \omega \boldsymbol{x}_t - s \nabla_{\boldsymbol{x}_t} L, \quad (13)$$

where $\omega$ is weight decay factor and $s$ is the learning rate. The process optimizes the image such that the anomaly score of PaDiM is minimized. However, the noisy gradients $\nabla_{\boldsymbol{x}_t} L$ will corrupt the image after some iterations, introducing significant noises to the image. We propose to leverage the diffusion model to denoise the gradients for high-quality reconstruction.

We assume that the gradients to the input image follow a Gaussian distribution $\nabla_{\boldsymbol{x}_t} L \sim \mathcal{N}(\mathbf{0}, \beta_t \boldsymbol{I})$ with variance $\beta_t$. Our target is to denoise the gradients to generate high-quality anomaly-free images. Notice that if the weight decay factor is set to be $\omega = \sqrt{1-s^2}$, each optimization step becomes a diffusion step with the noise $\epsilon_t \sim \mathcal{N}(\mathbf{0}, s^2\beta_t \boldsymbol{I})$. We denote the new variance as $\hat{\beta}_t = s^2\beta_t$. Then we can safely use the diffusion model to denoise the intermediate images during the optimization. Since the variance of noise $\epsilon_t$ is relatively small, we denoise the image every $N_d$ optimization step. We demonstrate the sampling process with Algorithm 1.

We visualize the intermediate steps of our gradient denoising process in Fig 7. The anomalous pixels are gradually altered to transform the input image into a normal image. We compare the reconstruction results with other state-of-the-art reconstruction-based anomaly detection methods in Fig. 4. The reconstructed image from our gradient denoising process keeps high-frequency details and successfully removes the anomaly pixels.

# 4. Experiments

## 4.1. Dataset and Implementation Details

**MVTec-AD** We evaluate our proposed method on the MVTec-AD dataset [3], an industrial anomaly detection benchmark that comprises 15 categories, including ten object classes and five texture classes. Each class contains approximately 200 anomaly-free images for training and 100 images with anomalies for testing. The dataset provides pixel-level segmentation ground truth for evaluation. The MVTec-AD dataset contains various anomalies, making it a comprehensive and ideal benchmark for anomaly detection evaluation.

**Metrics** We assess our pixel-level anomaly segmentation performance with two commonly used threshold-independent metrics: the Area Under the Receiver Operating Characteristic curve (AUROC), and Per-Region-Overlap (PRO) [3]. While AUROC equally measures performance for each pixel, it tends to favor larger area anomalies. To correctly assess the performance on both large and small area anomalies, we also evaluate our method with the PRO. To compute PRO, the area coverage ratios of each connected component are averaged for the same false positive rate. By repeatedly computing the values for the false positive rate from 0 to 0.3, we get a curve, and the normalized integral of this curve is the PRO-score. Unlike AUROC, the PRO metric equally measures the performance for large and small anomalies, which makes it a balanced evaluation metric for industrial anomaly detection.

**Implementation details** We train our diffusion model separately for each category of MVTec-AD[3]. We adopt the UNet network design with attention modules from improved diffusion [23]. Please refer to the supplementary for network details. The timestep for the diffusion process is set to be 1000 for training and 250 for reverse sampling. The diffusion model is trained for 10,000 iterations with batch size 2 on a single GPU for all the experiments. We adopt AdamW [21] as the optimizer with an annealing learning rate starting at 0.0001. We also adopt the exponential-moving-average (EMA) during evaluation and reconstruction. For the unified model, we train a single diffusion model on all the categories of MVTec-AD for 20,000 iterations. The class label is provided to the UNet [24] of the diffusion models for image reconstruction.

We resize the image to $(256, 256)$ and train the model with a 5 degree random rotation augmentation. For the pre-trained deep networks, we choose EfficientNet [30] pre-trained on ImageNet [9]. We set the ensembling factor $\alpha$ in 11 to $0.5$ for all categories for the same-hyperparameter setting. Our best results are achieved with ensemble factors adjusted for each category. We select three timesteps $T = \{5, 50, 100\}$ during the forward diffusion process to get three different noise scales. The anomaly scores predicted are averaged as the final output. We set the learning rate for the gradient denoising process to be $0.02$. The image is denoised once every $N_d = 5$ iteration.

## 4.2. Quantitative Results

We compare our anomaly localization results with CutPaste [19], Spade [6], PaDiM [8], DRAEM [40], CFlow [13], and UniAD [36]. We evaluate the results with two localization metrics: pixel-wise AUROC and Per-Region-Overlap (PRO) [3]. We present the results on the MVTec-AD benchmark in Tab. 1. Our model with the same hyperparameters for all categories boosts the PRO by $0.7\%$, and our best model with adjusted hyperparameters for each category improves the PRO by $1.1\%$, compared with previous state-of-the-art reconstruction-based methods. We show results on BTAD [22] in the supplementary.

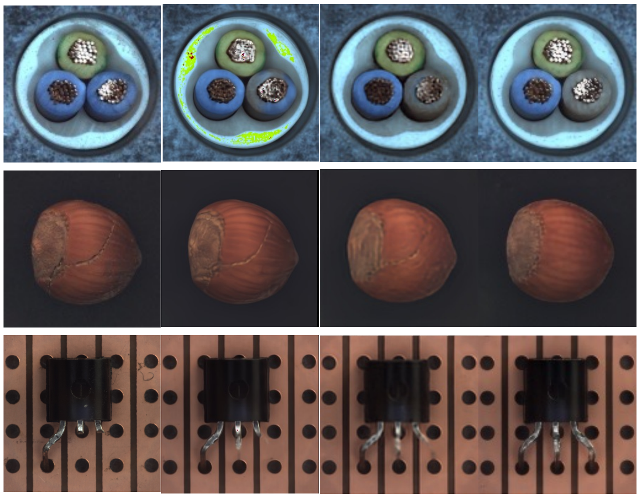
**Robustness.** Our study demonstrates the effectiveness of incorporating random noise to enhance the robustness of anomaly localization. We observed that the previous state-of-the-art reconstruction method DRAEM [40] uses pseudo anomalies that are unsuitable for detecting anomalies that vary from the pseudo data, particularly in the cable, pill, and transistor classes. The UniAD's transformer-based AutoEncoder [36] performs poorly on the metal nut, tile, and wood classes. In contrast, our denoising model learns the normal data distribution for anomaly detection, which is robust for all the categories.

## 4.3. Qualitative results of localization.

Figure 6 shows the anomaly localization results on MVTec-AD [3]. The first and fifth columns are images with anomalies from MVTec-AD. The columns from left to right

| Class | Embedding-based | | | Reconstruction-based | | | |
|---|---|---|---|---|---|---|---|
| | Spade [6] | PaDiM [8] | CFlow [13] | DRAEM[40] | UniAD [36] | Ours | Ours* |
| bottle | ( 98.4, 95.5) | (98.5, 95.3) | (99.0, 96.8) | (99.1, **97.2**) | (98.0, 93.8) | (97.7,95.0) | (97.7,95.0) |
| cable | ( 97.2, 90.9) | (98.1, 91.1) | (97.7, 93.5) | (94.7, 76.0) | (97.2, 86.3) | (95.2,88.7) | (95.6,**89.5**) |
| capsule | ( 99.0, 93.7) | (98.8, 92.3) | (99.0, 93.4) | (94.3, **91.7**) | (98.7, 90.8) | (98.0,90.1) | (97.5,91.4) |
| carpet | ( 97.5, 94.7) | (98.9, 94.5) | (99.3, 97.7) | (95.5, 92.9) | (98.4, 94.5) | (98.9,95.8) | (98.9,**95.8**) |
| grid | ( 93.7, 86.7) | (96.1, 90.5) | (99.0, 96.1) | (99.7, 98.4) | (97.5, 92.6) | (99.1,98.1) | (99.1,**98.4**) |
| hazelnut | ( 99.1, 95.4) | (98.4, 84.0) | (98.9, 96.7) | (99.7, **98.1**) | (98.2, 93.0) | (97.7,89.5) | (97.3,91.1) |
| leather | ( 97.6, 97.2) | (99.2, 97.9) | (99.7, 99.4) | (98.6, 98.0) | (98.7, 97.2) | (99.5,99.1) | (99.5,**99.1**) |
| metal nut | ( 98.1, 94.4) | (98.0, 92.9) | (98.6, 91.7) | (99.5, **94.1**) | (94.9, 87.1) | (96.8,93.0) | (96.8,93.0) |
| pill | ( 96.5, 94.6) | (97.0, 95.3) | (99.0, 95.4) | (97.6, 88.9) | (96.2, **95.3**) | (92.5,94.5) | (92.5,94.5) |
| screw | ( 98.9, 96.0) | (98.7, 94.6) | (98.9, 95.3) | (97.6, **98.2**) | (98.9, 95.3) | (99.0,95.6) | (99.0,95.6) |
| tile | ( 87.4, 75.9) | (94.3, 93.7) | (98.0, 94.3) | (99.2, **98.9**) | (92.0, 79.6) | (92.1,95.1) | (92.1,95.1) |
| toothbrush | ( 97.9, 93.5) | (98.7, 94.3) | (98.9, 95.1) | (98.1, 90.3) | (98.3, 88.2) | (98.9,94.7) | (98.6,**95.7**) |
| transistor | ( 94.1, 87.4) | (97.9, 91.4) | (98.0, 81.4) | (90.9, 81.6) | (97.9, **93.9**) | (92.6,89.7) | (93.1,90.1) |
| wood | ( 88.5, 87.4) | (95.7, 89.3) | (96.7, 95.8) | (96.4, **94.6**) | (93.0, 86.0) | (94.7,92.9) | (94.5,93.0) |
| zipper | ( 96.5, 92.6) | (98.5, 95.0) | (99.0, 96.6) | (98.8, **96.3**) | (97.7, 93.2) | (97.6,93.6) | (97.6,93.6) |
| average | ( 96.0, 91.7) | (97.8, 92.8) | (98.6, 94.6) | (97.3,93.0) | (97.0, 91.1) | (96.7,93.7) | (96.7,**94.1**) |

Table 1: Compare with the state-of-the-art anomaly detection approaches on MVTec-AD [3] dataset. We compare anomaly localization performance with pixel-wise AUROC and PRO metrics, denoted as (AUROC, PRO) in the table. We highlight the best PRO scores among all the reconstruction-based methods. We denote the result in Red if the result underperforms other reconstruction-based methods by a large margin. We show our results with the same hyperparameters for all categories, denoted as Ours, and different hyperparameters adjusted for each category, denoted as Ours*.



(a) Input (b) DRAEM (c) UniAD (d) Ours

Figure 4: Comparisons of the reconstruction results on class cable and hazelnut of MVTec-AD. The anomaly types are color-swap, crack, and cut-lead for the three categories.

represent ground truth, our pixel-level anomaly predictions, our feature-level anomaly predictions, and the final results visualized on the original images. We show that our denoising model is capable of precise boundary estimation of anomalies.

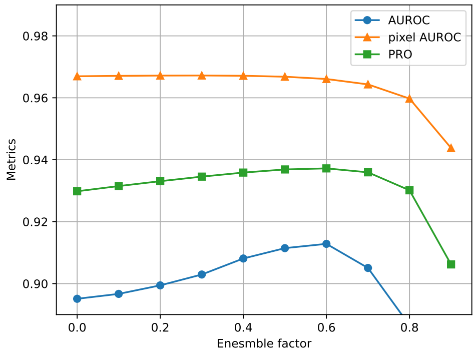


Figure 5: The effects of different ensembling factor $\alpha$.

## 4.4. Ablation study

**Ensemble factor.** We combine the pixel-level and feature-level 11 predictions to get the final anomaly score. We conduct experiments to verify the effects of the ensembling weights. As Fig. 5 shows, we choose the best ensemble factor $\alpha = 0.5$. The PRO metric is improved by $0.7\%$ compared with only using the feature-level anomaly score.

**Pretrained feature extractor.** We observe that the deep features extracted by ResNet [14] and WideResNet [39] are of high dimensional, which brings great difficulties for our denoising model to reconstruct the features. Instead, we

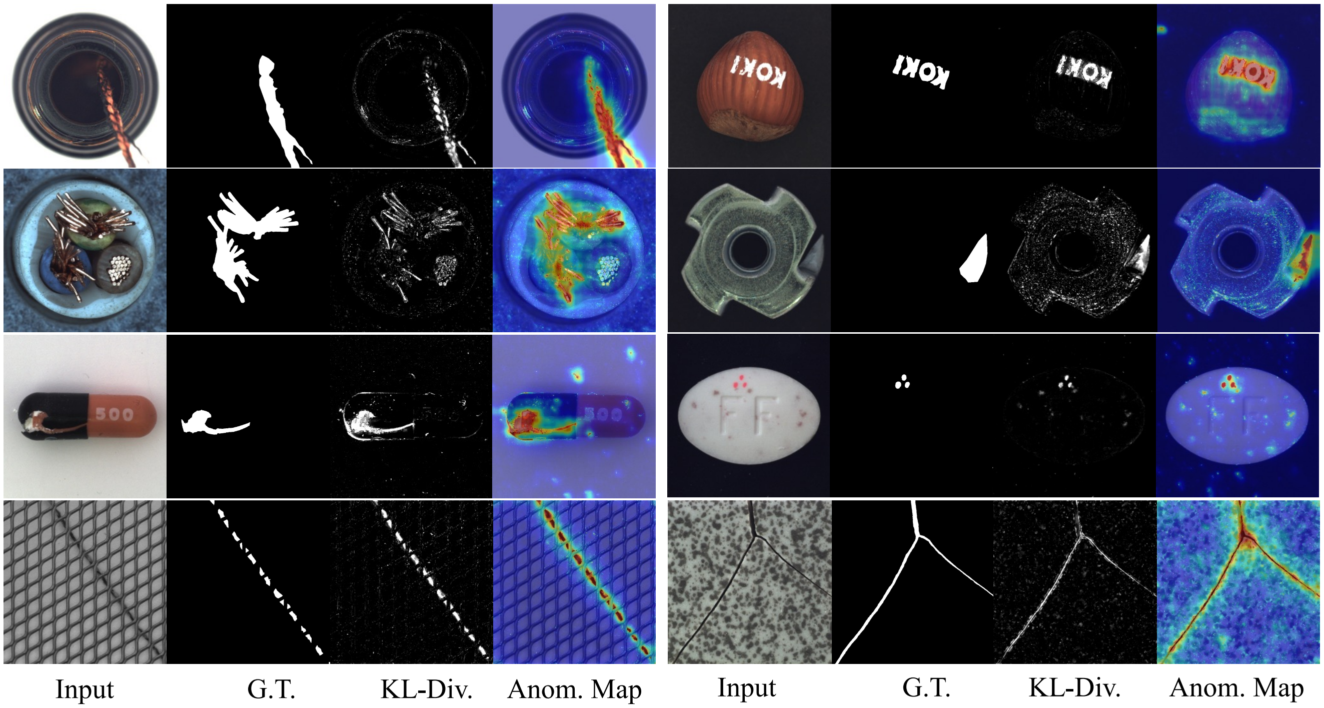


Figure 6: Visualization of anomaly localization on MVTec-AD [3]. The columns from left to right present anomalous images, the corresponding ground truths, our pixel-level predictions based on KL divergence, and the visualization of the final anomaly map on the original image. Our denoising diffusion models produce boundary-aware pixel-level anomaly scores to improve the quality of final anomaly scores.

adopt EfficientNet [30] as our feature extractor, where the dimension of features is relatively small. We select the intermediate feature map of stride $(2, 4, 8, 16)$ with dimension $(24, 32, 56, 160)$, resize to $64 \times 64$ and concatenate them together into a tensor $\boldsymbol{f} \in \mathbb{R}^{272\times64\times64}$. We conduct experiments to verify the resolution of the concatenated features; see Tab. 2. Increasing resolution from $16 \times 16$ to $64 \times 64$ increases the PRO by $2.5\%$. Combining EfficentNet features of stride $(2, 4, 8, 16)$ improves the PRO by $3.5\%$.

**Unified model.** Since our denoising diffusion model is capable of modeling complex real industrial data distribution, we conduct experiments to use a single model for anomaly localization for all categories of MVTec-AD, see Tab. 3. The performance of DRAEM [40] and PaDiM [8] drops greatly for the unified setting, with more than $5\%$ degradation in the PRO metric. In contrast, our model still achieves $93.0\%$ in PRO, with less than $1.1\%$ performance drop.

### 4.5. Qualitative results of reconstruction.

We show in Fig. 7 that our denoising gradient process can smoothly transform an anomalous image into a normal one under the guidance of a pre-trained feature extractor. The left two columns are ground truth and input anomalous

| Resize | Stride | | | | AUROC | PRO |
|---|---|---|---|---|---|---|
| | 16 | 8 | 4 | 2 | | |
| 64 | ✓ | | | | 94.2 | 90.2 |
| | ✓ | ✓ | | | 95.8 | 91.9 |
| | ✓ | ✓ | ✓ | | 96.4 | 93.2 |
| | ✓ | ✓ | ✓ | ✓ | **96.7** | **93.7** |
| 32 | ✓ | ✓ | ✓ | ✓ | 96.6 | 92.4 |
| 16 | ✓ | ✓ | ✓ | ✓ | 95.5 | 91.2 |

Table 2: Ablation study of feature level reconstruction. Experiments are conducted with features from different layers of EfficientNet [30], resized to different resolutions.

| Method | Base | | Unified | |
|---|---|---|---|---|
| | AUROC | PRO | AUROC | PRO |
| PaDiM[8] | **97.8** | 92.8 | 90.5 | 85.3 |
| DRAEM[40] | 97.3 | 93.0 | 89.4 | 82.2 |
| UniAD[36] | 96.6 | - | **97.0** | 91.1 |
| Ours | 96.7 | **94.1** | 96.0 | **93.0** |

Table 3: Comparison of a single unified model for anomaly localization of all the categories on MVTec-AD.

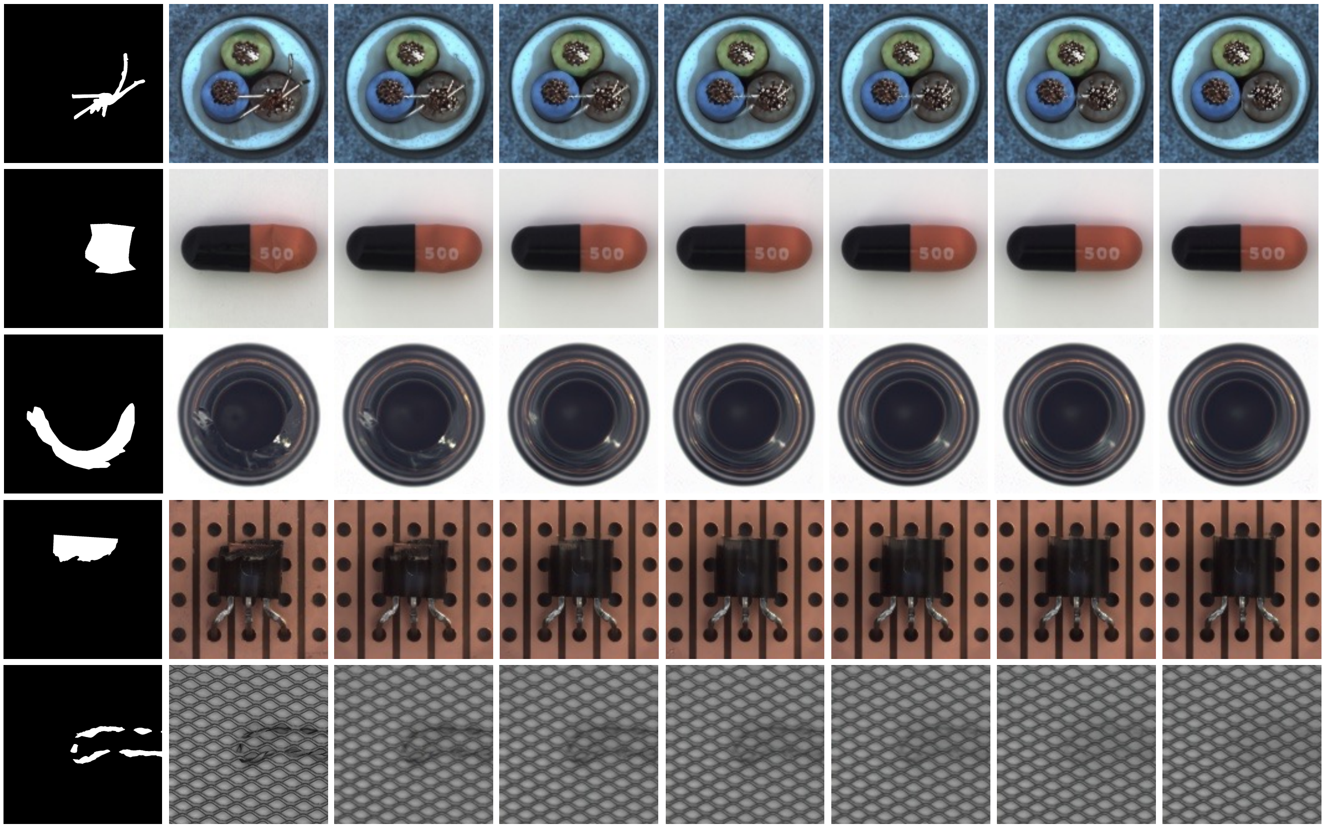


Figure 7: Visualization of the intermediate steps of denoising gradient process on MVTecAD. The 1st and 2nd columns are the ground truth of anomalies and the image to be reconstructed. The last column is the reconstructed results from the anomalous image. The intermediate results of the denoising gradient process are shown in the middle columns. We show that our reconstruction process can generate high-quality anomaly-free images and keep high-frequency details of normal pixels.

image, followed by the intermediate results from our gradient denoising process. The anomalous pixels are gradually removed by the gradients and diffusion steps. We show that the proposed reconstruction process can generate normal images with strong correspondence to the original anomalous image, while keeping most visual appearances of normal regions unchanged.

We compare the reconstruction results with previous anomaly detection methods, DRAEM [40] and UniAD [36]. As Fig. 4 shows, the DRAEM generates artifacts and fails to remove the anomalous pixels. The appearance of normal regions is changed and blurred by UniAD since it reconstructs the results with a VAE decoder. Our gradient noising process greatly improves the reconstruction results for both anomalous and normal pixels.

## 5. Conclusions

In this work, we propose a denoising diffusion model to boost the performance of reconstruction-based anomaly localization. Our model combines pixel-level and feature-level reconstruction errors as the anomaly score. We use the KL divergence from the diffusion model to produce boundary-aware results for better localization. Moreover, our model can reconstruct anomalous images to a high-quality normal image by denoising the gradients from a pre-trained deep feature extractor, surpassing the previous reconstruction results by a large margin. We also demonstrate that our reconstruction-based denoising diffusion model is robust to various anomaly types and can be extended as a unified anomaly detector for all categories.

**Discussions** Our approach addresses the anomaly localization problem from a denoising perspective. The MVTec-AD dataset contains many noises in the background regions, which are easily detected as anomalous by our denoising model, causing a $3\%$ performance drop on the image-level AUROC metric.

**Acknowledgement** The work was supported in part by the Research Grants Council under the Areas of Excellence scheme grant AoE/E-601/22-R and Shenzhen Science and Technology Program KQTD20210811090149095.

## References

[1] Samet Akcay, Amir Atapour-Abarghouei, and Toby P Breckon. Ganomaly: Semi-supervised anomaly detection via adversarial training. In *Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part III 14*, pages 622–637. Springer, 2019.

[2] Christoph Baur, Benedikt Wiestler, Shadi Albarqouni, and Nassir Navab. Deep autoencoding models for unsupervised anomaly segmentation in brain mr images. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 4th International Workshop, BrainLes 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers, Part I 4*, pages 161–169. Springer, 2019.

[3] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten 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*, pages 9592–9600, 2019.

[4] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Uninformed students: Student-teacher anomaly detection with discriminative latent embeddings. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4183–4192, 2020.

[5] Paul Bergmann, Sindy Löwe, Michael Fauser, David Sattlegger, and Carsten Steger. Improving unsupervised defect segmentation by applying structural similarity to autoencoders. *arXiv preprint arXiv:1807.02011*, 2018.

[6] Niv Cohen and Yedid Hoshen. Sub-image anomaly detection with deep pyramid correspondences. *arXiv preprint arXiv:2005.02357*, 2020.

[7] Anne-Sophie Collin and Christophe De Vleeschouwer. Improved anomaly detection by training an autoencoder with skip connections on images corrupted with stain-shaped noise. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 7915–7922. IEEE, 2021.

[8] Thomas Defard, Aleksandr Setkov, Angelique Loesch, and Romaric Audigier. Padim: a patch distribution modeling framework for anomaly detection and localization. In *Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10–15, 2021, Proceedings, Part IV*, pages 475–489. Springer, 2021.

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[10] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021.

[11] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*, 2016.

[12] Nanqing Dong and Eric P Xing. Few-shot semantic segmentation with prototype learning. In *BMVC*, volume 3, 2018.

[13] Denis Gudovskiy, Shun Ishizaka, and Kazuki Kozuka. Cflow-ad: Real-time unsupervised anomaly detection with localization via conditional normalizing flows. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 98–107, 2022.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[15] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

[16] Pavel Izmailov, Polina Kirichenko, Marc Finzi, and Andrew Gordon Wilson. Semi-supervised learning with normalizing flows. In *International Conference on Machine Learning*, pages 4615–4630. PMLR, 2020.

[17] Antanas Kascenas, Nicolas Pugeault, and Alison Q O'Neil. Denoising autoencoders for unsupervised anomaly detection in brain mri. In *International Conference on Medical Imaging with Deep Learning*, pages 653–664. PMLR, 2022.

[18] Yunseung Lee and Pilsung Kang. Anovit: Unsupervised anomaly detection and localization with vision transformer-based encoder-decoder. *IEEE Access*, 10:46717–46724, 2022.

[19] Chun-Liang Li, Kihyuk Sohn, Jinsung Yoon, and Tomas Pfister. Cutpaste: Self-supervised learning for anomaly detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9664–9674, 2021.

[20] Wenqian Liu, Runze Li, Meng Zheng, Srikrishna Karanam, Ziyan Wu, Bir Bhanu, Richard J Radke, and Octavia Camps. Towards visually explaining variational autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8642–8651, 2020.

[21] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[22] Pankaj Mishra, Riccardo Verk, Daniele Fornasier, Claudio Piciarelli, and Gian Luca Foresti. Vt-adl: A vision transformer network for image anomaly detection and localization. In *2021 IEEE 30th International Symposium on Industrial Electronics (ISIE)*, pages 01–06. IEEE, 2021.

[23] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pages 8162–8171. PMLR, 2021.

[24] Olaf Ronneberger, Philipp Fischer, and Thomas 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*, pages 234–241. Springer, 2015.

[25] Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter Gehler. Towards total recall in industrial anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14318–14328, 2022.

[26] Lukas Ruff, Robert Vandermeulen, Nico Goernitz, Lucas Deecke, Shoaib Ahmed Siddiqui, Alexander Binder, Emmanuel Müller, and Marius Kloft. Deep one-class classifica-

tion. In *International conference on machine learning*, pages 4393–4402. PMLR, 2018.

[27] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015.

[28] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

[29] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

[30] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.

[31] David MJ Tax and Robert PW Duin. Support vector data description. *Machine learning*, 54:45–66, 2004.

[32] Zhuotao Tian, Pengguang Chen, Xin Lai, Li Jiang, Shu Liu, Hengshuang Zhao, Bei Yu, Ming-Chang Yang, and Jiaya Jia. Adaptive perspective distillation for semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2):1372–1387, 2022.

[33] Kaixin Wang, Jun Hao Liew, Yingtian Zou, Daquan Zhou, and Jiashi Feng. Panet: Few-shot image semantic segmentation with prototype alignment. In *proceedings of the IEEE/CVF international conference on computer vision*, pages 9197–9206, 2019.

[34] Yizhou Wang, Dongliang Guo, Sheng Li, and Yun Fu. Towards explainable visual anomaly detection. *arXiv preprint arXiv:2302.06670*, 2023.

[35] Julian Wyatt, Adam Leach, Sebastian M Schmon, and Chris 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*, pages 650–656, 2022.

[36] Zhiyuan You, Lei Cui, Yujun Shen, Kai Yang, Xin Lu, Yu Zheng, and Xinyi Le. A unified model for multi-class anomaly detection. *arXiv preprint arXiv:2206.03687*, 2022.

[37] Zhiyuan You, Kai Yang, Wenhan Luo, Lei Cui, Yu Zheng, and Xinyi Le. Adtr: Anomaly detection transformer with feature reconstruction. *arXiv preprint arXiv:2209.01816*, 2022.

[38] Jiawei Yu, Ye Zheng, Xiang Wang, Wei Li, Yushuang Wu, Rui Zhao, and Liwei Wu. Fastflow: Unsupervised anomaly detection and localization via 2d normalizing flows. *arXiv preprint arXiv:2111.07677*, 2021.

[39] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.

[40] Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj. Draem-a discriminatively trained reconstruction embedding for surface anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8330–8339, 2021.