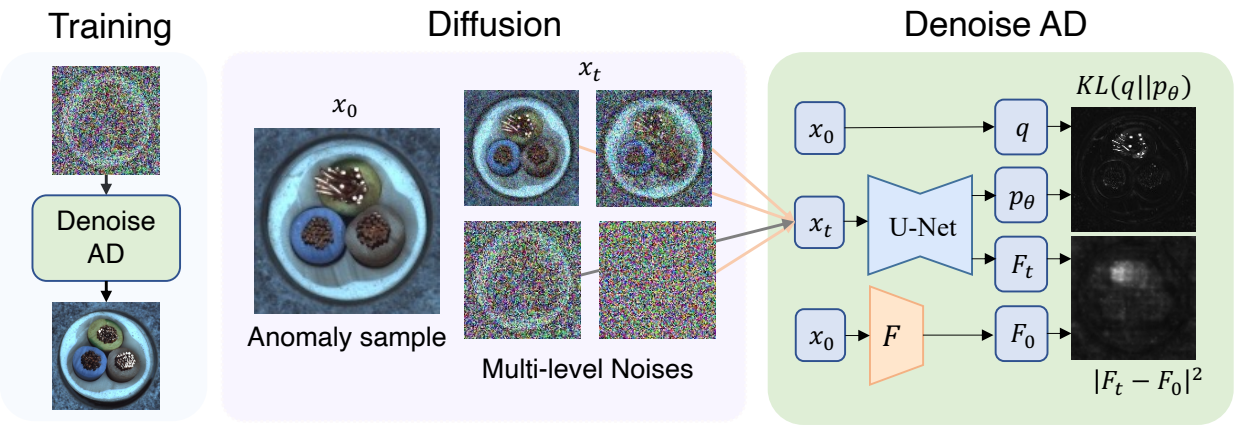

Figure 2: Our denoising diffusion model is trained with only anomaly-free images. During inference, noises of different scales are added to the anomaly sample. With large enough noises, the anomalous pixels become indistinguishable from the normal pixels and easier for reconstruction. We take the KL-divergence 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 ϵ.

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

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