Correcting Deviations from Normality: A Reformulated Diffusion Model for Multi-Class Unsupervised Anomaly Detection

Farzad Beizaee[1,2*] Gregory A. Lodygensky[2] Christian Desrosiers[1] Jose Dolz[1]
[1]ÉTS Montreal [2]CHU-Sainte-Justine Montreal

Abstract

Recent advances in diffusion models have spurred research into their application for Reconstruction-based unsupervised anomaly detection. However, these methods may struggle with maintaining structural integrity and recovering the anomaly-free content of abnormal regions, especially in multi-class scenarios. Furthermore, diffusion models are inherently designed to generate images from pure noise and struggle to selectively alter anomalous regions of an image while preserving normal ones. This leads to potential degradation of normal regions during reconstruction, hampering the effectiveness of anomaly detection. This paper introduces a reformulation of the standard diffusion model geared toward selective region alteration, allowing the accurate identification of anomalies. By modeling anomalies as noise in the latent space, our proposed ***Deviation correction diffusion*** *(DeCo-Diff) model preserves the normal regions and encourages transformations exclusively on anomalous areas. This selective approach enhances the reconstruction quality, facilitating effective unsupervised detection and localization of anomaly regions. Comprehensive evaluations demonstrate the superiority of our method in accurately identifying and localizing anomalies in complex images, with pixel-level AUPRC improvements of 11-14% over state-of-the-art models on well-known anomaly detection datasets. The code is available at https://github.com/farzad-bz/DeCo-Diff*

1. Introduction

Unsupervised anomaly detection (UAD) is paramount to a wide span of computer vision problems across strategic and high-impact domains, such as industrial inspection [2, 16], video surveillance [19, 30], or medical imaging [24, 26]. The main goal of UAD is to identify corrupted images, as well as anomalous pixels within these images, by leveraging only *normal* images. This setting naturally arises in many scenarios, where normal samples are readily available but compiling a curated set of labeled abnormal images is costly, due to the complexity of the annotation process as well as the scarcity and high variability of potential abnormalities.

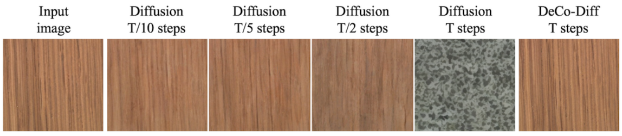

Figure 1. **Diffusion model reconstruction vs. DeCo-Diff.** Fine details and patterns of a normal image are changed during the standard forward-backward diffusion process: the "wood" image becomes a "Tile" sample when T steps are applied. In contrast, DeCo-Diff does not alter the details of the image using T correction steps, maintaining the appearance of the original input image.

The prevailing approach to address UAD frames this task as a *cold-start*[1] anomaly detection problem, which is typically approached through three primary categories of methods: reconstruction [20, 24, 31], synthesizing [13, 25, 34, 35] and embedding [2, 3, 22, 23] based approaches. While popular, single-class approaches hinder the scalability of these strategies, as the amount of storage and training time increases with the number of categories. Thus, there is a real need for novel methods that can accommodate the multi-class scenario in a robust and efficient manner.

Under the unified scenario, however, the distribution of normal data becomes more complex. As a result, the success of this task hinges on robust models capable of effectively learning the joint distribution across diverse types of objects. Diffusion models [27], and more particularly Denoising Diffusion Probabilistic Models (DDPM) [10], have emerged as strong candidates for this endeavor. Indeed, diffusion models have demonstrated strong potential in reconstructing the normal counterpart of an image to localize anomalous regions [4, 38].

Diffusion models, though highly effective at generating high-quality samples, are designed to generate images from pure noise. Therefore, their direct application to modify

*Corresponding author: farzad.beizaee.1@ens.etsmtl.ca

[1]Also commonly known as one-class classification (OCC).

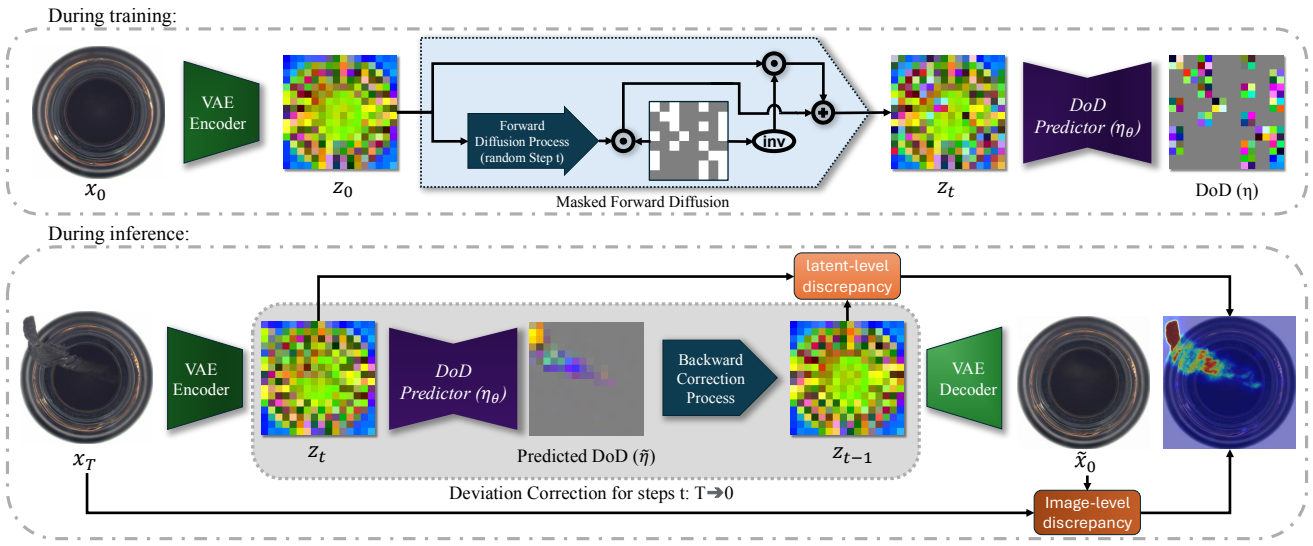

Figure 2. **Overview of the proposed method.** During **training** (*top*), normal images are partially diffused using random masks and randomly sampled time-steps ([1,T]). Then, our DeCo-Diff model is trained to predict the direction of deviation from the input image. At **inference** (*bottom*) starting from time-step T for the target images, DeCo-Diff progressively corrects the deviation from normality.

the input images necessitates a forward diffusion process followed by a backward denoising process. However, this strategy imposes several limitations. First of all, directly using DDPMs in the multi-class context may result in misclassifying generated images due to the loss of their original category information [8], when a high number of diffusion steps is applied. On the other hand, if we use lower diffusion steps, the *"identity shortcut"* issue will emerge, where the model tends to simply denoise the input regardless of whether the content is normal or anomalous, thus preserving the anomalous regions[33]. A further problem arises when forward and reverse diffusion steps are applied indiscriminately across the entire image. In such cases, the model may struggle to fully recover the normal patterns leading to false identification of anomalies due to discrepancies between the input image and its reconstruction. This problem is more significant when the input image involves random patterns and textures, such as "tile". Lastly, the model's ability to accurately reconstruct the normal appearance of anomalous regions improves by leveraging the contextual information from surrounding normal regions. However, diffusion models introduce noise across the entire image, leading to the loss of crucial details required for reconstructing the normal counterpart of the image. These limitations are depicted in Fig. 1.

To address these limitations, we propose a reformulation of the standard diffusion models in such a way that it corrects the deviation from normality. This new formulation better suits our goal of reconstructing abnormal regions into their normal counterpart, while preserving the fine details of normal areas. Specifically, it enables us to directly use the target anomalous image in the backward correction process during inference, which eliminates the need for forward diffusion process, prevents the degradation of informative information, and enables selective correction of abnormal regions.The proposed model also introduces a random masking strategy for the added noise, which brings two important benefits. First, it leaves portions of the image untouched during the forward process, thus making the correction process conditioned on those normal untouched regions. Secondly, during the backward correction process, using our reformulation, the model learns to selectively alter abnormal regions of the image while keeping normal areas unchanged. This leads to a more precise anomaly detection and localization.

Our **key contributions** can be summarized as follows:

- We propose a novel reformulation of diffusion models that learns to correct deviation from the learned distribution of normality to its normal counterparts, rather than generating samples through denoising steps.
- We introduce a random masking strategy into the forward diffusion process, which conditions the deviation correction of abnormal regions to surrounding areas while preserving the fine details of normal regions.
- Furthermore, we reformulated the DDIM sampling [28] to accommodate the deviation correction approach presented, enabling a faster, stable, and deterministic reverse process for efficient sampling.
- Extensive results on popular anomaly detection benchmarks demonstrate the superiority of our approach across recent state-of-the-art methods and multiple metrics.

2. Related work

2.1. Unsupervised Anomaly detection

Unsupervised anomaly detection has been studied from multiple perspectives: ***1) Reconstruction-based approaches*** rely on the assumption that a model trained on normal data will fail to reconstruct anomalous regions, leveraging the differences between the input and its reconstructed image as an anomaly score. Generative Adversarial Networks [20, 24, 31], Variational Auto-Encoders [15], and Normalizing Flows [5], are typically used as the backbone for reconstruction networks in this task. ***2) Synthesizing-based approaches*** typically introduce synthetic anomalies in normal images used for training [13, 25, 34, 35]. For example, DRÆM [34] trains a network on synthetically-generated just-out-of-distribution patterns, whereas CutPaste [13] introduces anomalies via a simple, yet efficient strategy based on cutting-pasting image patches at random locations on a normal image. ***3) Embedding-based approaches*** focus on embedding the normal features into a compressed space, relying on the assumption that anomalous features are far from the normal clusters in this space. These methods employ networks that are pre-trained on large-scale datasets like ImageNet for feature extraction [2, 3, 22, 23]. For instance, PaDiM [2] resorts to a multivariate Gaussian to model the distribution of normal patch features at each position of the image, and then measures the normality score using the Mahalanobis distance. Similarly, PatchCore [22] employs a core set to encode the features of normal patches, and finds anomalies at test time by computing the distance from a new patch's embedding and its nearest element in the core set. More recently, [14] proposed using the hyperbolic space to measure the distance between feature representations. Despite the progress made by these approaches in the single-class setting, their performance in the unified scenario remains unsatisfactory.

2.2. Multi-class unsupervised anomaly detection

The direct application of above-mentioned methods to the multi-class problem often leads to suboptimal performance. Moreover, since each category needs a separate trained model, the computational burden of these methods quickly explodes as the number of classes increases.

Tackling this more realistic scenario, multi-class unsupervised anomaly detection (Mc-UAD) approaches have recently gained traction within the community [3, 6–8, 16, 33, 36, 39]. UniAD [33] proposes a series of modifications to accommodate reconstruction-based networks for the challenging task of Mc-UAD. Observing the importance of the query embedding module in transformers to model the normal distribution, a query decoder is integrated into each layer, instead of only in the first layer as in vanilla transformers. To alleviate the "*identity shortcut*" problem in transformers, authors also introduce a neighbor mask attention module preventing tokens to copy themselves via self-attention. Last, a feature jittering strategy is employed to help the model recover normal features from noisy ones. Rd4AD [3] leverages a reverse distillation approach, where a student network learns to restore the multi-scale representations of a teacher given the teacher one-class embeddings. MambaAD [7] proposed a pyramidal autencoder framework to reconstruct multi-scale features using recently proposed Vision Mamba networks [40]. MoEAD [17] introduced a Mixture-of-Experts architecture to transform single-class models into a unified model. DiAD [8] leveraged diffusion models for multi-class AD by integrating a semantic-guided network that helps preserve semantic information in the reconstruction of a Latent Diffusion Model. More Recently, GLAD [32] proposed combining the global and local adaptive mechanisms to improve the reconstruction performance of diffusion models. However, diffusion-based methods indiscriminately apply noise to the entire image and lack an explicit mechanism for learning how to reconstruct abnormal regions from their normal surrounding regions.

3. Preliminaries

Denoising Diffusion Probabilistic Models (DDPM) are based on the idea of progressively perturbing data samples into noise via a forward process, and reversing this process to generate new data samples. This section introduces the fundamental concepts and notations required for understanding and applying diffusion models introduced in [10] and [28].

3.1. Forward Diffusion Process

Let $\mathcal{X} = \{\boldsymbol{x}^{(i)}\}_{i=1}^{N}$ denote the data samples (*i.e.*, images), where $\boldsymbol{x}^{(i)} \in \mathbb{R}^{H\times W\times C}$ is the *i*-th image, with H, W, and C representing its height, width and number of channels, respectively. The forward Markov diffusion process consists in gradually corrupting data samples by adding Gaussian noise for t time-steps where $t \in [1, T]$. Following [10], the forward noising process can be characterized as follows:

$$q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) = \mathcal{N}(\boldsymbol{x}_t; \sqrt{1-\beta_t}\boldsymbol{x}_t, \beta_t\mathbf{I}). \quad (1)$$

Here, β_i is a noise scheduling parameter that controls the amount of noise added at each step, such that $\boldsymbol{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ for a sufficiently large T. The marginal distribution at the moment t can then be explicitly defined from $\boldsymbol{x}_0$ as:

$$q(\boldsymbol{x}_t|\boldsymbol{x}_0) = \mathcal{N}(\boldsymbol{x}_t; \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0, (1-\bar{\alpha}_t)\mathbf{I}), \quad (2)$$

where $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{i=1}^{T} \alpha_i$.

3.2. Reverse Diffusion Process

In the reverse process, which iteratively removes noise through a series of denoising steps to generate new samples from Gaussian noise, the prior probability at each step

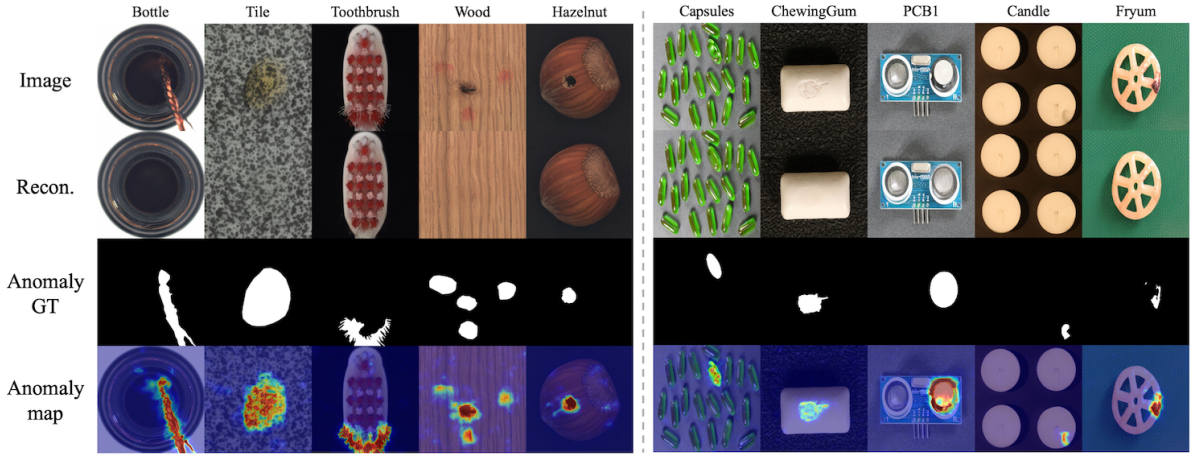

Figure 3. **Qualitative results.** From *top* to *bottom*: the original input image (with anomalies), DeCo-Diff reconstruction, the ground truth mask, and the predicted anomaly mask. Examples are depicted for two datasets (MVTec-AD on the left side and VisA on the right side) and across multiple anomalies with diverse complexity.

can be modeled as a Gaussian distribution:

$$p(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t) = \mathcal{N}(\boldsymbol{x}_{t-1}; \mu_\theta(\boldsymbol{x}_t, t), \beta_t \mathbf{I}), \tag{3}$$

in which β_t is fixed for each t, and the mean functions $\mu_\theta(\boldsymbol{x}_t, t)$ parameterized by θ are trainable:

$$\mu_\theta(\boldsymbol{x}_t, t) = \frac{1}{\sqrt{\bar{\alpha}_t}}\Big(\boldsymbol{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t)\Big). \tag{4}$$

This reverse process is trained with the variational lower bound of the log-likelihood of $\boldsymbol{x}_0$, *i.e.*:

$$\mathcal{L}(\theta) = -\,p(\boldsymbol{x}_0|\boldsymbol{x}_1) + \sum_t \mathcal{D}_{KL}\big(q^*(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0) \,\|\, p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)\big). \tag{5}$$

By applying the reparameterization trick (*i.e.*, parameterizing the noise function $\boldsymbol{\epsilon}_\theta$ as a neural network), the model can be trained using a simple mean-square error loss between the ground truth sampled Gaussian noise and the predicted noise:

$$\min_\theta \; \mathbb{E}_{\boldsymbol{x}_0 \sim q(\boldsymbol{x}_0), \boldsymbol{\epsilon}, t}\left[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta\,(\boldsymbol{x}_t, t)\|_2^2\right], \tag{6}$$

with $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Thus, starting from a noise vector $\boldsymbol{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ sampling is achieved by iteratively applying the learned reverse process to gradually reconstruct a clean sample $\boldsymbol{x}_0$.

4. Method

We resort to diffusion models to characterize the distribution of normal images, and reconstruct the normal counterparts of target images to detect and localize anomalies. However, standard diffusion models are inherently designed to generate new samples starting from pure noise and not altering selective regions of an input image. Moreover, by applying the forward and then reverse diffusion processes to reconstruct the normal counterpart of an anomalous image, depending on time-step t, the information in the normal regions can be partially degraded or lost. This degradation leads to a suboptimal reconstruction of normal areas, which is not ideal for unsupervised anomaly detection as input images and their reconstructed counterparts are compared to detect abnormal regions.

To address this issue, we propose a modified diffusion model (DeCo-Diff) that selectively alters only the anomalous regions of an image, while leaving the normal areas intact. This approach allows for preserving the normal parts of anomalous images while altering abnormal regions, based on the learned distribution of normal images. In this section, we detail the different components of our approach, whose overall pipeline is depicted in Fig. 2.

4.1. Modeling anomalies as noise in latent space

We employ a pre-trained Variational Auto-Encoder (VAE) [21] to project images into a latent space where the diffusion process is performed. We chose to work in this latent space for five key reasons: *i)* anomalies in the image can be effectively interpreted as noise within the latent space, which aligns well with the operational framework of diffusion models; it also *ii)* alleviates the problem of "identity shortcut" which causes anomalies to be preserved in the reconstruction; *iii)* increases computational efficiency; *iv)* enhances stability, especially when training data is limited; and *v)* improves the quality of generated samples.

Denoting the encoder and decoder networks of the VAE as ϕ_E and ϕ_D, an input image $\boldsymbol{x}^{(i)}$ is projected into a latent space representation $\boldsymbol{z}^{(i)} = \phi_E(\boldsymbol{x}^{(i)})$, where $\boldsymbol{z}^{(i)} \in$

$\mathbb{R}^{H' \times W' \times C'}$. It is important to note that, due to the unsupervised nature of this method, we only have access to normal images during the training phase.

Limitations of standard diffusion models. There are certain limitations to adding noise to the whole input image (or its latent representation) as in standard diffusion models. First, the added noise also affects normal regions. Consequently, during the forward diffusion process, normal regions may experience a partial or complete loss of information, making it difficult to fully recover the input image and potentially leading to their misinterpretation as anomalous areas when compared to the input image. Furthermore, abnormal regions should be reconstructed with respect to the surrounding normal areas. However, if normal patches are altered due to the forward diffusion process, they cannot be fully exploited to reconstruct abnormal patches, a critical step for anomaly detection.

To alleviate the aforementioned limitations, we propose to integrate random masking into the forward diffusion process. Given the latent features of an input normal sample $\boldsymbol{z}_0 \sim p(\boldsymbol{z}_0)$, we spatially divide $\boldsymbol{z}_0$ into non-overlapping noisy $\boldsymbol{z}_0^n$ (non-masked) and visible $\boldsymbol{z}_0^v$ patches (masked) using a random mask with a random masking ratio (r_{mask}). Afterward, during the forward process, only the noisy patches in $\boldsymbol{z}_0^n$ are gradually diffused, while the visible patches $\boldsymbol{z}_0^v$ remain unchanged:

$$q(\boldsymbol{z}_t^k|\boldsymbol{z}_0^k) = \begin{cases} \mathcal{N}(\boldsymbol{z}_t^n; \sqrt{\bar{\alpha}_t}\boldsymbol{z}_0^n, (1-\bar{\alpha}_t)\mathbf{I}), & \textit{if } k = n \\ \boldsymbol{z}_0^v, & \textit{otherwise.} \end{cases} \tag{7}$$

This way, visible patches in the image can represent normal regions that have been unaltered during the forward process, while noisy patches could mimic anomalous areas.

4.2. Deviation instead of Diffusion

While we motivated the use of a masking strategy (Eq. (7)) during the forward diffusion process, conditioning the reverse process and training objective on the masks used for the forward pass is suboptimal and impractical. First, accessing to prior knowledge on the location of abnormal regions during test time is unrealistic. Secondly, the denoising operation in Eq. (3) is not applicable to *non-noisy* patches. Thus, we need to integrate a mechanism that optimizes the training objective and applies the reverse process to both noisy and visible patches.

To achieve this, we reformulate the standard diffusion model to better suit our needs. Let us first consider the noisy patches. Based on Eq. (1), for noisy patches in the latent space we have:

$$\boldsymbol{z}_t^n = \sqrt{1-\beta_t}\boldsymbol{z}_{t-1}^n + \sqrt{\beta_t}\boldsymbol{\epsilon} \tag{8}$$

Based on Eq. (2), $\boldsymbol{z}_t^n$ could be explicitly obtained with

$$\boldsymbol{z}_t^n = \sqrt{\bar{\alpha}_t}\boldsymbol{z}_0^n + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon} \tag{9}$$

We can then rewrite Eq. (9) as:

$$\begin{aligned} \boldsymbol{z}_t^n &= (1-(1-\sqrt{\bar{\alpha}_t}))\boldsymbol{z}_0^n + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon} \\ &= \boldsymbol{z}_0^n + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon} - (1-\sqrt{\bar{\alpha}_t})\boldsymbol{z}_0^n \\ &= \boldsymbol{z}_0^n + \underbrace{\sqrt{1-\bar{\alpha}_t}\left(\boldsymbol{\epsilon} - \frac{1-\sqrt{\bar{\alpha}_t}}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{z}_0^n\right)}_{\textit{“Deviation from Normality”}} \end{aligned} \tag{10}$$

In the above equation, the second term measures the deviation of $\boldsymbol{z}_t^n$ from $\boldsymbol{z}_0^n$, which represents a normal image. In summary, we have that:

$$\begin{aligned} \boldsymbol{z}_t^n &= \boldsymbol{z}_0^n + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\eta} \\ \boldsymbol{\eta} &= \left(\boldsymbol{\epsilon} - \frac{1-\sqrt{\bar{\alpha}_t}}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{z}_0^n\right), \end{aligned} \tag{11}$$

where $\boldsymbol{\eta}$ is a term based on the time-step t, $\boldsymbol{z}_0$, and $\boldsymbol{\epsilon}$, which points to the “*Direction of Deviation*” (DoD). One interesting characteristic of this approach is that the patches will remain untouched in the forward process if $\boldsymbol{\eta} = \mathbf{0}$.

The reverse process aims to correct the deviation from a normal image by progressively removing noise through a learned denoising function $p(\boldsymbol{z}_{t-1}|\boldsymbol{z}_t)$. This could be achieved by training a neural network to closely approximate the true reverse path by predicting the DoD at each time-step. The main objective can thus be defined as:

$$\min_{\theta} \; \mathbb{E}_{\boldsymbol{z}_0 \sim q(\boldsymbol{z}_0), \boldsymbol{\epsilon}, t}\left[\|\boldsymbol{\eta} - \boldsymbol{\eta}_\theta(\boldsymbol{z}_t, t)\|_2^2\right]. \tag{12}$$

This approach encourages the model to predict the DoD from normality at each time-step, facilitating accurate correction in the reverse process. As a result, our DoD predictor ($\boldsymbol{\eta}_\theta$) learns a robust denoising trajectory that can correct the noisy, abnormal patches with respect to the surrounding normal ones, while keeping the normal patches untouched by simply predicting zero as their DoD. Furthermore, to expose the model to more structured anomalies during training, we also propose to incorporate a strategy based on *patch shuffling*. Since the latent space follows a normal distribution, for a portion of patches (specified with $r_{shuffle}$), instead of adding Gaussian noise, we can replace them with other patches taken from images in the same batch. This strategy could ultimately result in better localization of large and structured anomalies.

4.3. Anomaly detection

In order to detect anomalies at inference time, we first need to adapt the sampling of DDPM (Eq. (3)) to fit the proposed configuration for correcting the deviation from normality. However, before describing this step, it is worth mentioning that sampling with DDPMs requires many reverse sampling steps until $t = 0$, leading to a significant computational burden. In addition, DDPMs introduce noise over the

Table 1. **Quantitatve evaluation on MVTec-AD.** Image and Pixel-level results on *multi-class* anomaly detection. Best method (per metric) is highlighted in **blue**, whereas **red** is used to denote the second-best approach. Differences with the second (or best) approach in (gray).

Method	Image-level AUROC	Image-level AUPRC	Image-level $f1_{max}$	Pixel-level AUROC	Pixel-level AUPRC	Pixel-level $f1_{max}$	Pixel-level AUPRO
RD4AD *CVPR'22*	94.6	96.5	95.2	96.1	48.6	53.8	91.1
UniAD *NeurIPS'22*	96.5	98.8	96.2	96.8	43.4	49.5	90.7
SimpleNet *CVPR'23*	95.3	98.4	95.8	96.9	45.9	49.7	86.5
DeSTSeg *CVPR'23*	89.2	95.5	91.6	93.1	54.3	50.9	64.8
DiAD *AAAI'24*	97.2	99.0	96.5	96.8	52.6	55.5	90.7
MoEAD *ECCV'24*	98.0	99.3	97.5	96.9	49.8	43.5	91.0
GLAD *ECCV'24*	97.5	99.1	96.6	97.4	**60.8**	**60.7**	93.0
MambaAD *NeurIPS'24*	**98.6**	**99.6**	**97.8**	**97.7**	56.3	59.2	**93.1**
DeCo-Diff (*Ours*)	**99.3**$_{+0.7}$	**99.8**$_{+0.2}$	**98.5**$_{+0.7}$	**98.4**$_{+0.7}$	**74.9**$_{+14.1}$	**69.7**$_{+9.0}$	**94.9**$_{+1.8}$

whole image at each step t, which goes against our objective of keeping normal patches unchanged. Consequently, we resort to the Denoising Diffusion Implicit Model (DDIM) [28] during inference, which modifies the sampling process by making it deterministic instead of stochastic. To correct the deviation using DDIM sampling, with a trained model $\boldsymbol{\eta}_\theta$, the predicted normal image ($\tilde{\boldsymbol{z}}_0$) at each time-step t is:

$$\tilde{\boldsymbol{z}}_0 = \boldsymbol{z}_t - \sqrt{1-\bar{\alpha}_t}\boldsymbol{\eta}_\theta(\boldsymbol{z}_t, t). \tag{13}$$

Therefore, the reverse process becomes:

$$\tilde{\boldsymbol{z}}_{t-1} = \tilde{\boldsymbol{z}}_0 + \sqrt{1-\bar{\alpha}_{t-1}}\ \underbrace{\boldsymbol{\eta}_\theta(\boldsymbol{z}_t, t)}_{\text{“Direction of deviation”}} \tag{14}$$

In the above equation, $\tilde{\boldsymbol{z}}_0$ is the predicted $\boldsymbol{z}_0$, and $\boldsymbol{\eta}_\theta(\boldsymbol{z}_t, t)$ the DoD. Note that Eq. (14) enables a deterministic trajectory from $\boldsymbol{z}_\tau$ to $\boldsymbol{z}_0$, which accommodates our setting since it avoids introducing any noise. It is noteworthy to mention that a key advantage of DDIM over DDPM is that, by eliminating the noise at each step and making the denoising process more direct, high-quality samples can be generated in fewer steps. This results in a faster convergence to normal samples.

Intuitively, as our DeCo-Diff model has been solely trained using normal images, when an anomalous image is provided the anomalous regions will fall outside the learned distribution. Consequently, the model will consider them as non-masked noisy patches, which will be denoised during the reverse process. On the other hand, if the patches do not correspond to anomalous regions, the model will consider them as masked normal patches and they will remain unchanged during the reconstruction. Furthermore, in contrast to standard diffusion processes that add noise to the whole image, the proposed model allows leveraging surrounding normal information when reconstructing abnormal regions.

Leveraging multi-scale information. Pixel-level reconstruction discrepancy is widely used in reconstruction-based UAD approaches mainly due to this ability to produce detailed anomaly maps. Nevertheless, some anomalous regions might not be revealed using this approach, for example, if they only present small differences in color compared to the normal ones. On the other hand, reconstruction discrepancy measured in the latent space can capture more subtle anomalies, however, it yields coarser anomaly maps due to the reduced spatial dimension of the latent space.

Based on these observations, we propose a multi-scale strategy where discrepancies at both pixel-space and latent-space are leveraged jointly to localize potential anomalies. More concretely, in addition to the reconstructed “normal” feature embedding $\tilde{\boldsymbol{z}}_0$, we also generate a reconstructed normal sample in the image-space as $\tilde{\boldsymbol{x}}_0 = \phi_D(\tilde{\boldsymbol{z}}_0)$, where ϕ_D is the pre-trained VAE decoder. The predicted anomaly map can be thus obtained using the geometric mean of latent-level and pixel-level discrepancies as:

$$\boldsymbol{a} = \sqrt{\frac{1}{\gamma_l \cdot \gamma_p} \min(\|\tilde{\boldsymbol{z}}_0 - \boldsymbol{z}_0\|, \gamma_l) \cdot \min(\|\tilde{\boldsymbol{x}}_0 - \boldsymbol{x}_0\|, \gamma_p)}. \tag{15}$$

Here, γ_l and γ_p respectively denote the latent-level and pixel-level thresholds, which prevent assigning excessive weight to patches with significant deviations in either the image or the latent space. For instance, if a black patch is reconstructed as white, this does not necessarily indicate that the patch is more anomalous compared to a patch reconstructed as red.

5. Experiments

5.1. Setting

Datasets. We evaluate our method on two well-known anomaly detection datasets. **MVTec-AD [1]** simulates real-world industrial production scenarios, filling the gap in unsupervised anomaly detection. It consists of 5 types of textures and 10 types of objects, in 5,354 high-resolution images from different domains. **VisA [41]** consists of 10,821 high-resolution images, with 78 types of anomalies. It comprises 12 subsets corresponding to distinct objects categorized into three object types: Complex structure, Mul-

Table 2. **Quantitatve evaluation on VisA.** Image and Pixel-level results on *multi-class* anomaly detection. The best method (per metric) is highlighted in **blue**, whereas **red** is used to denote the second-best approach. Differences with second (or best) approach in (gray).

Method	Image-level			Pixel-level			
	AUROC	AUPRC	$f1_{max}$	AUROC	AUPRC	$f1_{max}$	AUPRO
RD4AD *CVPR'22*	92.4	92.4	89.6	98.1	38.0	42.6	91.8
UniAD *NeurIPS'22*	88.8	90.8	85.8	98.3	33.7	39.0	85.5
SimpleNet *CVPR'23*	87.2	87.0	81.8	96.8	34.7	37.8	81.4
DeSTSeg *CVPR'23*	88.9	89.0	85.2	96.1	**39.6**	43.4	67.4
DiAD *AAAI'24*	86.8	88.3	85.1	96.0	26.1	33.0	75.2
MoEAD *ECCV'24*	93.0	**95.1**	**89.8**	**98.7**	36.6	41.0	88.6
GLAD *ECCV'24*	91.8	92.9	88.0	97.4	34.6	40.3	**92.0**
MambaAD *NeurIPS'24*	**94.3**	94.5	89.4	98.5	39.4	**44.0**	91.0
DeCo-Diff (*Ours*)	**96.4**$_{+2.1}$	**96.8**$_{+1.7}$	**92.2**$_{2.4}$	**98.5**$_{-0.2}$	**51.3**$_{+11.7}$	**51.2**$_{+7.2}$	**92.1**$_{+0.1}$

tiple instances, and Single instance. Note that ablations are performed on MVTec-AD. More information about the datasets is provided in Appendix A.

Evaluation Metrics. Following prior literature [7, 8], we use Area Under the Receiver Operating Characteristic Curve (AUROC), Area Under Precision-Recall Curve (AUPRC[2]), and f1-score-max ($f1_{max}$) to measure performance in both anomaly detection (i.e., *image-level*) and anomaly localization (i.e., *pixel-level*). Furthermore, Area Under Per-Region-Overlap (AUPRO) is employed for pixel-level anomaly localization.

Baselines. We compare our approach to eight recently-proposed approaches for multi-class UAD: RD4AD [3], UniAD [33], SimpleNet [16], DeSTSeg [39], DiAD [8], MoEAD[17], GLAD [32], and MambaAD [7].

Implementation Details. To map data into the latent space using a pre-trained VAE, we used a pre-trained perceptual compression model [21], which consists of an auto-encoder trained by a combination of a perceptual loss [37] and a patch-based adversarial objective that down-samples the image by a factor 8. Similar to VAEs[12], a regularization loss measuring the KL divergence between latent features and $\mathcal{N}(\mathbf{0}, \mathbf{I})$ is used to avoid the latent space having an arbitrarily high variance.

For our model $\boldsymbol{\eta}_\theta$ predicting the *direction of deviation* (DoD), following [21], we employed an attention UNet architecture with timestep embedding and a squared-cosine-beta scheduler [18]. At each iteration and for each training sample, we first randomly select a timestep t from a uniform distribution between 1 and T (where T=10). Then, we sample a masking ratio (r_{mask}) from a uniform distribution $U[0, 0.7]$ and a latent patch size from $\{1, 2, 4, 8\}$ corresponding to pixel-wise patch sizes of $\{8\times8, 16\times16, 32\times32, 64\times64\}$. Using these sampled values, we generate a random mask with the specified ratio. Moreover, in our implementation, the replacement ratio ($r_{shuffle}$) is sampled in each iteration from $U[0, 0.3]$. Finally, $\boldsymbol{z}_t$ is obtained using Eq. (7), and serves as input to train our model.

[2]Also called Average Precision(AP) or AUPR in the literature

This model was trained 800 epochs for experiments on the MVTec dataset and 200 epochs for VisA, in both cases using a batch size of 128 and a single A6000 GPU. We used the AdamW optimizer with a Cosine Annealing scheduler with warm-up, with an initial learning rate set to 10^{-4}, decaying to a minimum of 10^{-5}. The VAE remained frozen throughout training.

5.2. Results

5.2.1. Quantitative Results

MVTec-AD: Tab. 1 reports the performance for all methods, where we can see that the proposed DeCo-Diff obtains the overall best performance across all metrics for both image-wise and pixel-wise multi-class anomaly detection. More concretely, DeCo-Diff achieves 0.7%, 0.2%, 0.7% improvements over the second best approach, i.e., MambaAD [7], for image-level AUROC, AUPRC, and $f1_{max}$, respectively. The boost in performance is even more notable at pixel-level, an arguably more complex scenario. In particular, our method yields improvements of 0.2%, 14.1%, 9.0%, and 1.8% over the existing state-of-the-art multi-class UAD models measured by pixel-level AUROC, AUPRC, $f1_{max}$, and AUPRO, respectively. ***VisA:*** Results in Tab. 2 confirm the trend observed in MVTec-AD, with our approach outperforming recent methods in both pixel-level and image-level metrics. Particularly, DeCo-Diff obtains around 2% performance gain over the second-best approach in image-level metrics, and exhibits significant pixel-level improvement over all methods in terms of AUPRC, $f1_{max}$, and AUPRO, yet ranking second for pixel-level AUROC. Detailed per-category results are provided in Appendix B.

5.2.2. Qualitative Results

The quantitative results presented above are supported by visual examples showcasing the effectiveness of our approach in identifying anomalous image regions. In Fig. 3, we depict anomalous samples from both datasets and their reconstruction using our model. As it can be observed, abnormal regions are properly modified during the deviation

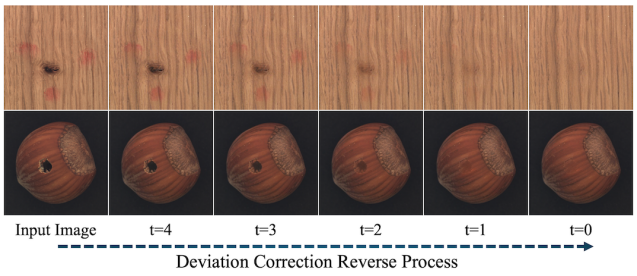

Figure 4. **Visualization of deviation correction through the reverse process steps.** Abnormal areas are progressively corrected, while fine normal details are preserved during the reverse process.

correction process and replaced by their normal counterpart. In contrast, the normal regions remained unchanged, our model preserving the fine details in these regions. This enables a precise detection and localization of abnormalities when contrasting the input image to its reconstructed version (Fig. 3, *last row*). Moreover, Fig. 4 visually depicts the deviation correction reverse process from DeCo-Diff, which showcases how abnormal areas are progressively removed, until a *cleaned* version of the input image is generated.

5.2.3. Ablation Studies

Effect of backward deviation correction strategy and number of time-steps. In this section, we explore different strategies for the deviation correction process during inference. In Eq. (14), we explained how the proposed deviation correction formulation can be used in the DDIM reverse sampling to progressively correct the abnormalities in each time-step t. An alternative strategy would be to directly replace $\boldsymbol{z}_{t-1}$ with $\tilde{\boldsymbol{z}}_0$ at each time-step. This approach allows the model to correct the input image based on its initial prediction at the first time-step, and then continue refining any remaining abnormalities in subsequent steps.

We evaluated the performance for different number of reverse steps using both approaches, where the first approach is referred to as $\tilde{\boldsymbol{z}}_{t-1}$ and the latter as $\tilde{\boldsymbol{z}}_0$ in Tab. 3. Note that, for a single reverse step, both approaches are identical and yield the same results. As reported in Tab. 3, the first strategy ($\tilde{\boldsymbol{z}}_{t-1}$) performs better in image-level metrics, while the second yields slightly better performance in terms of pixel-level metrics. Furthermore, the performance is generally consistent as the number of reverse time-steps increases, making a lower number of time-steps (but larger than 1) more favorable due to its decreased computational cost.

Effect of leveraging discrepancies at pixel, latent, and visual feature levels. In reconstruction-based anomaly detection, mapping the differences between inputs and their reconstructions to anomaly maps is essential. Thus, we now investigate several alternatives to the proposed strategy, presented in Eq. (15). First, we consider using only the difference in pixel values, i.e., $\Delta\boldsymbol{x} = |\boldsymbol{x}_T - \boldsymbol{x}_0|$, which are clipped and scaled by $\gamma_l = 0.4$, as: $\frac{1}{\gamma_l}\min(\Delta\boldsymbol{x}, \gamma_l)$. However, this option may not be optimal, as anomalies might appear with the same color as normal regions. Then, we resort to the difference in the latent space ($\Delta\boldsymbol{z} = |\boldsymbol{z}_T - \boldsymbol{z}_0|$), which is similarly clipped and scaled by the same value, and then resized to match the target image size. Additionally, we explored the impact of the arithmetic mean: $\frac{1}{2\gamma_l}\min(\Delta\boldsymbol{z}, \gamma_l) + \frac{1}{2\gamma_p}\min(\Delta\boldsymbol{x}, \gamma_p)$. Finally, we extracted the visual features from a pre-trained ResNet50 [9] for both the input images and their reconstructions and compared them using the cosine similarity to localize anomalies. The results for anomaly detection and localization using different levels of discrepancies are presented in Tab. 4. These results support our approach (i.e., Geometric mean, Eq. (15)), and demonstrate that combining pixel- and latent-space discrepancies with the geometric mean yields superior performance to considering these discrepancies separately.

Table 3. Impact of the correction strategy at each time-step and number of reverse steps.

Correction strategy	Reverse steps	Image-level AUROC / AUPRC / $f1_{mask}$	Pixel-level AUROC / AUPRC / $f1_{mask}$ / AUPRO
$\tilde{\boldsymbol{z}}_0$	1	98.4 / 99.3 / 97.9	98.4 / 72.9 / 68.8 / 94.3
$\tilde{\boldsymbol{z}}_0$	2	99.2 / 99.7 / 98.1	**98.4** / **75.1** / **69.9** / 94.7
$\tilde{\boldsymbol{z}}_0$	5	99.2 / 99.7 / 97.9	98.2 / 74.3 / 69.3 / 94.6
$\tilde{\boldsymbol{z}}_0$	10	98.9 / 99.5 / 97.6	97.8 / 73.6 / 68.8 / 94.2
$\tilde{\boldsymbol{z}}_{t-1}$	2	99.3 / 99.8 / 98.3	98.4 / 75.0 / 69.8 / 94.4
$\tilde{\boldsymbol{z}}_{t-1}$	5	**99.3** / **99.8** / **98.5**	98.4 / 74.9 / 69.7 / **94.9**
$\tilde{\boldsymbol{z}}_{t-1}$	10	99.3 / 99.8 / 98.5	98.4 / 74.9 / 69.8 / 95.0

Table 4. **Different levels of discrepancy.** Results on MVTec-AD.

Dissimilarity	Image-level AUROC / AUPRC / $f1_{mask}$	Pixel-level AUROC / AUPRC / $f1_{mask}$ / AUPRO
Pixel diff. ($\Delta\boldsymbol{x}$)	98.9 / 99.6 / 97.7	97.5 / 68.0 / 63.7 / 92.1
Latent diff. ($\Delta\boldsymbol{z}$)	99.2 / 99.7 / 98.3	98.2 / 70.5 / 67.5 / 93.5
Arithmetic ($\Delta\boldsymbol{x}$, $\Delta\boldsymbol{z}$)	**99.3** / **99.8** / **98.5**	98.2 / 73.5 / 68.8 / 94.1
Geometric ($\Delta\boldsymbol{x}$, $\Delta\boldsymbol{z}$)	**99.3** / **99.8** / **98.5**	**98.4** / **74.9** / **69.7** / **94.9**
Features (ResNet50)	98.5 / 99.3 / 97.3	97.4 / 66.7 / 64.6 / 91.7

6. Conclusions

We proposed a reformulation of standard diffusion models, specifically designed to modify abnormal regions in target images without affecting normal areas. By integrating masked strategy to the new formulation in the latent space, DeCo-Diff progressively detects and corrects the deviations from normality, enabling precise localization of abnormalities. Our comprehensive Quantitative and Qualitative results demonstrate the superiority of our model compared to state-of-art UAD methods in the unified multi-class context.

References

[1] 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. 6, 1

[2] Thomas Defard, Aleksandr Setkov, Angelique Loesch, and Romaric Audigier. Padim: a patch distribution modeling framework for anomaly detection and localization. In *International Conference on Pattern Recognition*, pages 475–489, 2021. 1, 3

[3] Hanqiu Deng and Xingyu Li. Anomaly detection via reverse distillation from one-class embedding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9737–9746, 2022. 1, 3, 7, 5

[4] Matic Fučka, Vitjan Zavrtanik, and Danijel Skočaj. Transfusion–a transparency-based diffusion model for anomaly detection. In *European conference on computer vision*, pages 91–108, 2024. 1

[5] 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. 3

[6] Jia Guo, Lize Jia, Weihang Zhang, Huiqi Li, et al. Recontrast: Domain-specific anomaly detection via contrastive reconstruction. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[7] Haoyang He, Yuhu Bai, Jiangning Zhang, Qingdong He, Hongxu Chen, Zhenye Gan, Chengjie Wang, Xiangtai Li, Guanzhong Tian, and Lei Xie. Mambaad: Exploring state space models for multi-class unsupervised anomaly detection. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 3, 7, 4, 5

[8] Haoyang He, Jiangning Zhang, Hongxu Chen, Xuhai Chen, Zhishan Li, Xu Chen, Yabiao Wang, Chengjie Wang, and Lei Xie. A diffusion-based framework for multi-class anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8472–8480, 2024. 2, 3, 7, 4, 5

[9] 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. 8

[10] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 1, 3

[11] Stepan Jezek, Martin Jonak, Radim Burget, Pavel Dvorak, and Milos Skotak. 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)*, pages 66–71. IEEE, 2021. 1, 5

[12] Diederik P Kingma. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 7

[13] 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. 1, 3

[14] Huimin Li, Zhentao Chen, Yunhao Xu, and Junlin Hu. Hyperbolic anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17511–17520, 2024. 3

[15] 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. 3

[16] Zhikang Liu, Yiming Zhou, Yuansheng Xu, and Zilei Wang. Simplenet: A simple network for image anomaly detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20402–20411, 2023. 1, 3, 7, 4, 5

[17] Shiyuan Meng, Wenchao Meng, Qihang Zhou, Shizhong Li, Weiye Hou, and Shibo He. Moead: A parameter-efficient model for multi-class anomaly detection. In *European Conference on Computer Vision*, 2024. 3, 7, 4

[18] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, pages 8162–8171. PMLR, 2021. 7

[19] Guansong Pang, Cheng Yan, Chunhua Shen, Anton van den Hengel, and Xiao Bai. Self-trained deep ordinal regression for end-to-end video anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12173–12182, 2020. 1

[20] Pramuditha Perera, Ramesh Nallapati, and Bing Xiang. Ocgan: One-class novelty detection using gans with constrained latent representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2898–2906, 2019. 1, 3

[21] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion mod-

els. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 4, 7

[22] 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. 1, 3

[23] Marco Rudolph, Tom Wehrbein, Bodo Rosenhahn, and Bastian Wandt. Fully convolutional cross-scale-flows for image-based defect detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1088–1097, 2022. 1, 3

[24] Thomas Schlegl, Philipp Seeböck, Sebastian M Waldstein, Ursula Schmidt-Erfurth, and Georg Langs. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In *International conference on information processing in medical imaging*, pages 146–157. Springer, 2017. 1, 3

[25] Hannah M Schlüter, Jeremy Tan, Benjamin Hou, and Bernhard Kainz. Natural synthetic anomalies for self-supervised anomaly detection and localization. In *European Conference on Computer Vision*, pages 474–489, 2022. 1, 3

[26] Julio Silva-Rodríguez, Valery Naranjo, and Jose Dolz. Constrained unsupervised anomaly segmentation. *Medical Image Analysis*, 80:102526, 2022. 1

[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. 1

[28] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021. 2, 3, 6

[29] Chengjie Wang, Wenbing Zhu, Bin-Bin Gao, Zhenye Gan, Jiangning Zhang, Zhihao Gu, Shuguang Qian, Mingang Chen, and Lizhuang Ma. Real-iad: A real-world multi-view dataset for benchmarking versatile industrial anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22883–22892, 2024. 1, 5

[30] Cheng Yan, Shiyu Zhang, Yang Liu, Guansong Pang, and Wenjun Wang. Feature prediction diffusion model for video anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5527–5537, 2023. 1

[31] Xudong Yan, Huaidong Zhang, Xuemiao Xu, Xiaowei Hu, and Pheng-Ann Heng. Learning semantic context from normal samples for unsupervised anomaly detection. In *Proceedings of the AAAI conference on artificial intelligence*, pages 3110–3118, 2021. 1, 3

[32] Hang Yao, Ming Liu, Haolin Wang, Zhicun Yin, Zifei Yan, Xiaopeng Hong, and Wangmeng Zuo. Glad: Towards better reconstruction with global and local adaptive diffusion models for unsupervised anomaly detection. In *European Conference on Computer Vision*, 2024. 3, 7, 4, 5

[33] Zhiyuan You, Lei Cui, Yujun Shen, Kai Yang, Xin Lu, Yu Zheng, and Xinyi Le. A unified model for multi-class anomaly detection. *Advances in Neural Information Processing Systems*, 35:4571–4584, 2022. 2, 3, 7, 4, 5

[34] 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. 1, 3

[35] Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj. Dsr–a dual subspace re-projection network for surface anomaly detection. In *European conference on computer vision*, pages 539–554, 2022. 1, 3

[36] Jiangning Zhang, Xuhai Chen, Yabiao Wang, Chengjie Wang, Yong Liu, Xiangtai Li, Ming-Hsuan Yang, and Dacheng Tao. Exploring plain vit reconstruction for multi-class unsupervised anomaly detection. *arXiv preprint arXiv:2312.07495*, 2023. 3

[37] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 7

[38] Xinyi Zhang, Naiqi Li, Jiawei Li, Tao Dai, Yong Jiang, and Shu-Tao Xia. Unsupervised surface anomaly detection with diffusion probabilistic model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6782–6791, 2023. 1

[39] Xuan Zhang, Shiyu Li, Xi Li, Ping Huang, Jiulong Shan, and Ting Chen. Destseg: Segmentation guided denoising student-teacher for anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3914–3923, 2023. 3, 7, 4, 5

[40] Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. Vision mamba: Efficient visual representation learning with bidirectional state space model. *arXiv preprint arXiv:2401.09417*, 2024. 3

[41] Yang Zou, Jongheon Jeong, Latha Pemula, Dongqing Zhang, and Onkar Dabeer. Spot-the-difference self-supervised pre-training for anomaly detection and segmentation. In *European Conference on Computer Vision*, pages 392–408. Springer, 2022. 6, 1