Enhancing Multi-Class Anomaly Detection through Diffusion Refinement with Dual Conditions

Anonymous CVPR submission

Paper ID 13725

Abstract

Anomaly detection, the technique of identifying abnormal samples using only normal samples, is a critical issue in industry. Existing one-model-per-category methods often struggle with limited generalization capabilities due to their focus on a single category, and can fail when encountering variations in product. Recent feature reconstruction methods, as representatives in one-model-all-categories schemes, face challenges including reconstructing anomalous samples and blurry reconstructions. In this paper, we creatively combine a diffusion model and a transformer for multi-class anomaly detection. This approach leverages diffusion to obtain high-frequency information for refinement, greatly alleviating the blurry reconstruction problem while maintaining the sampling efficiency of the reverse diffusion process. The task is transformed into image inpainting to disconnect the input-output correlation, thereby mitigating the "identical shortcuts" problem and avoiding the model from reconstructing anomalous samples. Besides, we introduce category-awareness using dual conditions to ensure the accuracy of prediction and reconstruction in the reverse diffusion process, preventing excessive deviation from the target category, thus effectively enabling multi-class anomaly detection. Futhermore, spatio-temporal fusion is also employed to fuse heatmaps predicted at different timesteps and scales, enhancing the performance of multi-class anomaly detection. Extensive experiments on benchmark datasets demonstrate the superior performance and exceptional multi-class anomaly detection capabilities of our proposed method compared to others.

1. Introduction

Anomaly detection has become an indispensable tool in a variety of fields, including industrial defect detection [2], medical image analysis [12], and video surveillance [47]. The technique is designed to identify abnormal images and locate abnormal regions. In industrial settings, anomaly detection is particularly challenging due to the scarcity of abnormal samples and the wide range of anomalies, which can range from subtle changes to large structural defects. Although existing methods [4, 9, 19, 22, 49, 54] have achieved acceptable performance, the practicality of their one-model–per-category scheme is debatable. When the deviation of the trained semantic category is large, these methods can easily fail and their generalization performance is severely limited [50]. In contrast, the one-model-all-categories scheme, which uses a unified framework to detect anomalies across various categories, can operate more effectively and reduce deployment costs in such scenarios.

As representatives of multi-class anomaly detection, reconstruction-based anomaly detection methods [1, 3, 20, 39, 53] have recently garnered increased interest. The generalization performance of such models allows them to be seamlessly ported into a unified model for multiple classes without compromising performance. However, these methods are not without their challenges. Firstly, they operate under the assumption that the model will only reconstruct normal samples after being trained on normal datasets, but the model has the potential to reconstruct anomalous samples, resulting in low reconstruction errors that fail to identify anomalous regions. Secondly, these methods involve processes like downsampling, which leads to the problem of blurred outputs and may cause the model to potentially overlook finer defects. This, in turn, can negatively impact the performance of anomaly detection.

To alleviate these problems, we propose a novel framework for multi-class anomaly detection named Diffusion Refinement with Dual Conditions (DRDC). Why choose diffusion? ❶ Several typical unsupervised anomaly detection methods using generative models, such as GANs [1, 13, 28, 53] and VAEs [29, 43], still suffer from potential resolution loss due to pooling and strided convolutions. Moreover, their performance is unsatisfactory, and the training process can become unstable on larger samples [6, 7, 26]. Since diffusion models offer excellent inductive biases for spatial data, the heavy spatial downsampling of related generative models is not required [15, 32]. Besides, the spatial dimensions of intermediate variables remain constant at each timestep, resulting in less loss of spatial information. ❷ You and Cui [50] provide insight into the nature of

"identical shortcuts"; that is the the characteristic of models to map inputs directly to outputs via shortcuts, resulting in the tendency to reconstruct anomalous samples when the inputs are abnormal. The "identity shortcut" problem can be mitigated, however, by inheriting the diffusion property that the back-end network uses to predict the noise.

How to exploit diffusion? ❶ Our method alleviates the blurry reconstruction problem by applying high-resolution refinement to the feature reconstruction-based model. The refinement is achieved by applying the diffusion model to an inpainting task that isolates the association with known and masked regions by making predictions, further preventing the model from mapping the input region directly to the output (i.e., potential leakage of "identical shortcuts") and thus mitigating the tendency to reconstruct anomalous samples. Moreover, due to the slow sampling speed of diffusion, direct reconstruction by diffusion either requires an unacceptable cost for inference time or a compromise in reconstruction accuracy. In contrast, we use the diffusion method to reconstruct the high-frequency part, which not only allows faster sampling, but also makes good use of the transformer-based underlying capability. ❷ The one-model–all-categories scheme poses a challenge when dealing with different anomaly definitions across categories, as what is considered anomalous in one category may be regarded as normal in another. Direct incorporation of the original diffusion can lead to uncontrollable randomness in the reconstructed images. While this doesn't significantly impact the one-model–per-category scheme, as we only incorporate samples from the same category for training, it can lead to the reconstruction of semantically different samples in the "one-model-all-categories" scenario, severely hampering performance. To achieve category-awareness to handle different anomaly definitions for multi-class anomaly detection, we introduce dual conditions to ensure that samples can be accurately reconstructed in the direction of the original semantic category. ❸ Moreover, we propose a spatio-temporal fusion module to perform a smoother fusion of heatmaps predicted at different timesteps and scales, which helps to mitigate the accumulative errors that can arise from relying solely on the final prediction or the bias of limited information at each timestep.

We conduct extensive experiments on the challenging MVTec-AD [2] and BTAD [25] to demonstrate the effectiveness of our method under the unified setting. Note that our DRDC achieves state-of-the-art 98.5 image-wise AUROC and 98.1 pixel-wise AUROC for multi-class anomaly detection on MVTec-AD. Our main contributions include:

- We determine the adverse effects of low spatial resolution from reconstruction methods. As a result, we propose a novel multi-class anomaly detection framework with diffusion refinement as a solution.
- Unlike the previous method of using diffusion in a trivial way, we avoid "identical shortcuts" by using diffusion for inpainting, while exploiting only the high-frequency part to refine the low-resolution heatmap, thus improving the sampling speed and taking advantage of diffusion model.
- To ensure accuracy and to accommodate the setting of multi-class anomaly detection, we introduce the technique of dual conditions and propose the module of spatio-temporal fusion, which enables a smooth fusion of heatmaps predicted at different timesteps and scales.
- Extensive experiments on MVTec-AD and BeanTechAD demonstrate the superiority of our framework over existing alternatives under the unified task setting.

2. Related Work

Anomaly Detection. Existing unsupervised anomaly detection techniques can be classified into classical methods [22, 38], distribution-based methods [9, 31], and reconstruction-based methods [3, 20, 51]. These one-model-per-category methods, however, are not well-suited when extending to multiple categories, as they hinge on a single hyper-sphere or a single distribution, which may not adequately capture the varying feature representation among categories. Recently, multi-class unified methods have received attention due to their practicality [16, 45, 50]. The most relevant UniAD [50] constructs a unified model for multiple categories via feature reconstruction, but its low spatial resolution (only 14×14 spatial resolution) of input features hinders its ability to precisely locate defects. Consequently, we introduce a diffusion-based refinement strategy for multi-class anomaly detection. This not only circumvents the issue of blurry reconstruction but also ensures superior anomaly detection performance, surpassing current state-of-the-art methods.

Diffusion Models. Lately, DDPMs have been in the spotlight for their ability on image synthesis [11], with superior mode coverage [48] over generative adversarial networks (GANs) and variational autoencoders (VAEs) [5]. Although there have been several endeavors in the anomaly detection domain [27, 44, 46], they primarily focus on medical image detection, which presents application scenarios vastly different from industrial anomaly detection. Alternatively, they have simply utilized diffusion models that offer mediocre performance and require a greater number of iteration steps. In contrast, we exploit only high-frequency information to increase the sampling speed while converting to inpainting tasks to avoid "identical shortcuts", thereby enhancing the usability of diffusion for anomaly detection.

3. Preliminaries

3.1. Base Model with Feature Reconstruction

For unsupervised multi-class anomaly detection, a base model is constructed with feature reconstruction, which dis-

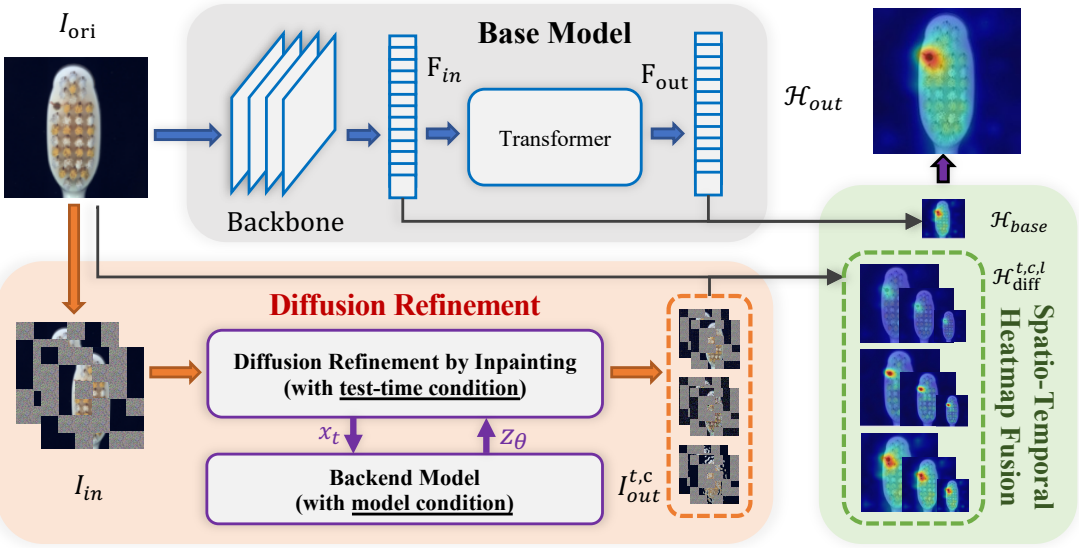

Figure 1. The overview of our proposed framework, including Base Model, Diffusion Refinement, and Spatio-Temporal Heatmap Fusion. The samples are first fed into a feature reconstruction network and a diffusion model. The feature reconstruction network is supported by a Transformer model, which produces a base heatmap, while the diffusion model is supported by a UNet model and the proposed dual conditions, which yields a high-resolution high-frequency correction heatmap. Finally, the spatio-temporal fusion module is exploited to obtain final high-resolution anomaly map.

criminates the anomaly via measuring the reconstruction error between the reconstructed sample and the original one. As in Fig. 1, the base model utilizes a transformer network to generate reconstructed feature for any input feature.

Formally, the base model is optimized by the Mean Squared Error (MSE) loss:

$$\mathcal{L}_{\text{feat}} = \frac{1}{H_{\text{feat}} \times W_{\text{feat}}} ||F_{\text{in}} - F_{\text{out}}||_2^2, \tag{1}$$

where, $F_{\text{in}} \in \mathbb{R}^{C_{\text{feat}} \times H_{\text{feat}} \times W_{\text{feat}}}$ is the input feature extracted from the original image by the pre-trained backbone network with fixed weights, and $F_{\text{out}} \in \mathbb{R}^{C_{\text{feat}} \times H_{\text{feat}} \times W_{\text{feat}}}$ is the recovered output feature generated by the transformer network. Note that C_{feat}, H_{feat} and W_{feat} represent the channel number, height, and width of the feature, respectively.

The anomaly result predicted by the base model is represented as a base heatmap $\mathcal{H}_{\text{base}} \in \mathbb{R}^{H_{\text{feat}} \times W_{\text{feat}}}$, which is the L2 norm of the reconstruction differences:

$$\mathcal{H}_{\text{base}} = ||F_{\text{in}} - F_{\text{out}}||_2. \tag{2}$$

3.2. Blurry Problem of Base Model

As previously stated, due to the limited spatial size of the features, with a resolution of merely 14×14, the final representation fails to accurately localize defects when the original resolution heatmap of 224×224 is obtained through direct upsampling, let alone for higher resolution. To illustrate this phenomenon, we downsample the ground-truth mask directly to a lower resolution, and then upsample it back to the original size. This process illustrates the significant performance loss that even the ground-truth mask would experience under such sampling treatment. For simplicity of illustration, we give the final pixel-wise AUROC averaged over 15 categories on MVTec-AD [2].

Table 1. Anomaly localization results of scaled ground-truth mask with AUROC metric on MVTec-AD.

Scaling Size	1	1/2	1/4	1/8	1/16
Pixel-wise AUROC	100.00	99.99	99.93	99.27	96.05

As revealed in Tab. 1, the impact on pixel-wise AUROC is not significant when the size is reduced by only half. However, when reduced to 1/16, the performance of pixel-wise AUROC declines relatively more. Note that these results cannot be directly regarded as the upper bound, since the data-type outputted by the model differs from that of the ground-truth mask (floating-point scores contain more information vs. binary masks can lose thin defects during downsampling). However, such results tell the limitations of small-size reconstruction. For this reason, we will perform a refinement to improve the anomaly map.

3.3. Diffusion Models

As outlined in [15, 40], diffusion models consist of a forward diffusion process and a reverse denoising process. The forward process is a Markov process that transforms an image x_0 into white Gaussian noise $x_0 \sim \mathcal{N}(0, 1)$ over a series of T timesteps. Each step is defined as,

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t\mathbb{I}), \tag{3}$$

where $\mathbb{I}$ is a identity matrix. Sample x_t is obtained by adding independent and identically distributed (i.i.d.) Gaussian noise with variance β_t at timestep t and scaling the previous sample x_{t-1} with $\sqrt{1-\beta_t}$ according to a variance schedule. By utilizing the independence property of the noise added at each step, as described in Eq. (3), we can calculate the total noise variance as $\bar{\alpha}_t = \prod_{s=1}^{t}(1-\beta_s)$, thereby marginalizing the forward process at each step:

$$q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)\mathbb{I}). \tag{4}$$

It allows us to efficiently sample pairs of training data to train a reverse transition step.

The diffusion model is trained to reverse the process in Eq. (3). The reverse process is modeled by a network that predicts μ_θ and Σ_θ of a Gaussian distribution,

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta, \Sigma_\theta). \tag{5}$$

The learning objective for the model is derived by considering the variational lower bound, as noted in [15],

$$\begin{aligned}
&\mathcal{L}_{\text{VLB}} = \mathcal{L}_T + \mathcal{L}_{T-1} + ... + \mathcal{L}_0, \\
&\mathcal{L}_t = \begin{cases} D_{KL}(q(x_T|x_0)||p(x_T)), & t = T, \\ D_{KL}(q(x_{t-1}|x_t, x_0)||p_\theta(x_{t-1}||x_t)), & 0 < t < T, \\ -\log p_\theta(x_0|x_1), & t = 0. \end{cases}
\end{aligned} \tag{6}$$

The subdomain $0 < t < T$ is crucial for training the network to perform a single reverse diffusion step. Additionally, it enables us to express the objective in closed form, as $q(x_{t-1}|x_t, x_0)$ is also Gaussian distribution.

According to [15], the most effective way to parameterize the model is to predict the cumulative noise ϵ_0 that is added to the current intermediate image x_t. Thus, we arrive at the following expression for the predicted mean μ_θ, where z_θ is the predicted noise:

$$\mu_\theta = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}z_\theta\right). \tag{7}$$

From the subdomain $0 < t < T$ of Eq. (6), the following simplied training objective can be derived.

$$\mathcal{L}_{\text{simple}} = \mathbb{E}\left[||\epsilon - z_\theta||_2^2\right]. \tag{8}$$

According to [41], the reverse diffusion process can be re-derived in the form of a non-Markovian chain, where the variance $\Sigma_\theta(x_t, t)$ in Eq. (5) can be removed and the number of required sampling steps can be reduced. As a result, we adopt this approach to improves the efficiency of the inference process.

4. Methodology

4.1. Our DRDC Framework

As depicted in Fig. 1, our framework primarily comprises three components: a base model, a diffusion refinement model, and a spatio-temporal fusion module. The base model is a transformer-like network, which reconstructs the input feature to produce a low-resolution base heatmap. In contrast, the diffusion refinement model produces multiple heatmaps that contain high-frequency information, taking into account different timesteps and inpainting specifications. In addition, our diffusion is guided by a dual conditional strategy, ensuring the model's category awareness and enhancing the accuracy of the reverse diffusion process. Lastly, we employ spatio-temporal fusion to fuse all the generated heatmaps from various timesteps, multiple scales, and different inpainting specifications.

4.2. Diffusion Refinement by Inpainting

Based on the discussion above, it's clear that while the base model can generate a acceptable heatmap, its limited resolution restricts its ability to accurately localize subtle defects, resulting in missed detections. In particular, categories such as screws and capsules, which contain thin strips of defects, are constrained by the low spatial resolution of the feature reconstruction. To tackle this issue, we employ a diffusion model to carry out image inpainting. This process involves recovering the masked region in the image by introducing noise and gradually reducing it through a reverse diffusion process. Notably, we perform inpainting at the same resolution as the original input image, which allows us to retain detailed spatial information. As the network is trained on normal samples, the recovered region can only be what it should look like under normal conditions.

For clarity, let's define $x_0 = I_{\text{in}}$, which serves as the input to this module. The input image x_0 is created by replacing a set of pixels with Gaussian noise. Specifically, we first divide the input image into $G = \frac{H_{\text{img}}}{c} \times \frac{W_{\text{img}}}{c}$ grids, where H_{img} and W_{img} are the image's height and width, and $c \in \{c_i\}_{i=1}^{n_c}$ is a factor of both H_{img} and W_{img}. These grids are then randomly and uniformly partitioned into n_s disjoint sets $\mathbb{S}_i$, each containing $\frac{G}{n_s}$ grid elements, with each element being a $c \times c$ pixel square. $\mathbb{M}_i \in \mathbb{M}$ is the corresponding binary mask that contains zeros in regions belonging to $\mathbb{S}_i$ and ones in other regions. During inference, $\mathbb{M}_i$ is used to replace the regions in the image that belong to $\mathbb{S}_i$ with Gaussian noise, transforming I_{ori} into I_{in}. The regions replaced with noise are gradually reconstructed back to $\tilde{x}_0$ after being processed forward as x_t using the forward diffusion process.

To leverage different information at different moments, we can apply a set of different timesteps $t \in \{t_i\}_{i=1}^{n_t}$. For each t, we can predict the masked region at timestep 0 using a reverse diffusion process as follows according to Eq. (4):

$$\tilde{x}_0^t = \frac{1}{\sqrt{\bar{\alpha}_t}}(x_t - \sqrt{1-\bar{\alpha}_t}z_\theta). \tag{9}$$

Combining the unknown part of each disjoint set $\mathbb{S}_i$ using its corresponding binary mask $\mathbb{M}_i$, we can get the pre-

dicted output $I_{\text{out}}^{t,c}$ of grid size c at timestep t.

$$I_{\text{out}}^{t,c} = \frac{1}{n_s} \sum_{\mathbb{M}_i \in \mathbb{M}} (1 - \mathbb{M}_i) \odot \tilde{x}_0^t. \tag{10}$$

As a result, heatmap $\mathcal{H}_{\text{diff}}^{t,c}$ can be obtained as follows:

$$\mathcal{H}_{\text{diff}}^{t,c} = ||I_{\text{ori}} - I_{\text{out}}^{t,c}||_2 \in \mathbb{R}^{H_{\text{img}} \times W_{\text{img}}}. \tag{11}$$

As per prior studies [8, 24], in the forward process, the original data progressively sheds information, with high-frequency details being lost before the low-frequency ones. Conversely, in the reverse process, information is gradually recovered from pure noise, with low-frequency details being obtained before the high-frequency ones. Drawing from these observations, we've designed our method to execute only a few timesteps $\{t\}$ during the early stages of the diffusion. This approach allows us to retain high-frequency information while also speeding up the reverse process.

4.3. Dual Conditioning on the Known Region

To perform multi-class unsupervised anomaly detection, it is necessary to accurately determine the category of a given sample. The base model encodes category information through an additional query embedding, while we utilize diffusion to implicitly encode categories as conditions without the need for an explicit classifier. The introduction of model conditioning and test-time conditioning ensures that the model does not misclassify similar categories, preventing it from incorrectly filling the missing region with the wrong semantic category.

Model Conditioning: Recall that the neural network model z_θ is trained to estimate ϵ for any given noisy image x_t. To improve the efficiency and accuracy of the neural network, we setup category-awareness into the model by adding conditions. Specifically, we condition the model on the input image I_{in} [with noise and with mask], allowing it to learn category-specific noise distributions and accurately reconstruct the target noise during the reverse diffusion process. As we aim to accelerate the inference process by limiting the number of timesteps required to obtain high-resolution information, it is important to ensure the accuracy of the predictions at each single timestep. Thus, the inclusion of conditions in the model is necessary. As a result, we extend Eq. (8) as follows,

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{t, x_0, I_{\text{in}}, \epsilon} \left[||\epsilon - z_\theta(x_t, I_{\text{in}}, t)||_2^2 \right], \tag{12}$$

where x_t is the sample at timestep t, I_{in} is our image condition. Note that we do not direct exploit our binary inpainting mask as in previous work [52, 55]. Instead, we fill the masked region of original images with Gaussian noise to meet the requirements of the diffusion model. The effectiveness for the model condition will be demonstrated in the ablation study section.

Test-time Conditioning: The condition at the test-time ensures that the model can guide the inpainting process of reverse diffusion, preventing excessive deviations from the original category. Drawing inspiration from [23], we denote the ground truth image as x, the known pixels as $\mathbb{M}_i \odot x$ and the unknown pixels as $(1 - \mathbb{M}_i) \odot x$.

Given that every reverse step in Eq. (5) from x_t to x_{t-1} depends solely on x_t, we can alter the known regions $\mathbb{M}_i \odot x_t$ and thereby insert information into the reverse process to control its noise reduction direction as long as we keep the correct properties of the corresponding distribution. Since the forward process in Eq. (3) is characterized by a Markov Chain of added Gaussian noise, we can sample the intermediate image x_{t-1} for the known region at any given time using Eq. (4),

$$x_{t-1}^{\text{known}} \sim \mathcal{N}(\sqrt{\bar{\alpha}_{t-1}} x_0, (1 - \bar{\alpha}_{t-1})\mathbb{I}). \tag{13}$$

This allows us to sample the known regions $\mathbb{M}_i \odot x_{t-1}$ at any timestep. Meanwhile, we can predict the intermediate image x_{t-1} for the unknown region using Eq. (5),

$$x_{t-1}^{\text{unknown}} \sim \mathcal{N}(\mu_\theta(x_t, I_{\text{in}}, t), \Sigma_\theta(x_t, I_{\text{in}}, t)). \tag{14}$$

Therefore, we achieve the following expression for one reverse step in our approach instead of Eq. (5),

$$x_{t-1} = \mathbb{M}_i \odot x_{t-1}^{\text{known}} + (1 - \mathbb{M}_i) \odot x_{t-1}^{\text{unknown}}. \tag{15}$$

In short, x_{t-1}^{known} is sampled using the known pixels in the given image $\mathbb{M}_i \odot x_0$, while x_{t-1}^{unknown} is sampled from the model given the previous iteration x_t. The results are then combined to the new sample x_{t-1} using the mask.

4.4. Spatio-temporal Heatmap Fusion

The primary purpose of the spatio-temporal fusion module is to fuse the low-resolution heatmap from the base model and high-resolution heatmaps from the diffusion model.

Recall that each step of the reverse diffusion process is able to predict a noise and to derive its corresponding predicted original image. However, each prediction corresponding to the current timestep leads to bias due to the limited information available. In contrast, using only the final predictions of the reverse diffusion leads to accumulative errors [14, 37]. In addition, several studies [4, 34, 36] have shown that performing fusion at different spatial scales can be helpful for anomaly detection. For the diffusion reconstructed heatmap, since anomalies tend to occupy larger spatially connected regions, the reconstruction error can be aggregated over a larger region for more accurate anomaly detection. All things considered, we propose a spatio-temporal fusion module to fuse various timesteps and spatial scales. For this purpose, we further extend Eq. (11) with a scaled $l \in \{l_i\}_{i=1}^{n_l}$. Thus, the anomaly score map for each scale l is calculated by downsampling the orginal image I_{ori}

Table 2. **Anomaly detection results with AUROC metric on MVTec-AD**. All methods are evaluated under the unified case. In the unified case, the learned model is applied to detect anomalies for all categories *without* fine-tuning.

Method	Object										Texture					Avg
	Bottle	Cable	Capsule	Hazelnut	MetalNut	Pill	Screw	Toothbrush	Transistor	Zipper	Carpet	Grid	Leather	Tile	Wood	
Student-Teacher [4]	84.0	60.0	57.6	95.8	62.7	56.1	66.9	57.8	61.0	78.6	86.6	69.2	97.2	93.7	90.6	74.5
PatchSVDD [49]	85.5	64.4	61.3	83.9	80.9	89.4	80.9	99.4	77.5	77.8	63.3	66.0	60.8	88.3	72.1	76.8
PaDiM [9]	97.9	70.9	73.4	85.5	88.0	68.8	56.9	95.3	86.6	79.7	93.8	73.9	99.9	93.3	98.4	84.2
CutPaste [19]	67.9	69.2	63.0	80.9	60.0	71.4	85.2	63.9	57.9	93.5	93.6	93.2	93.4	88.6	80.4	77.5
MKD [36]	98.7	78.2	68.3	97.1	64.9	79.7	75.6	75.3	73.4	87.4	69.8	83.8	93.6	89.5	93.4	81.9
DRÆM [54]	97.5	57.8	65.3	93.7	72.8	82.2	92.0	90.6	74.8	98.8	98.0	99.3	98.7	99.8	**99.8**	88.1
AST [35]	96.9	92.3	93.6	98.6	98.8	84.9	92.0	97.2	96.1	86.2	97.6	93.4	99.4	96.6	98.2	94.7
SimpleNet [21]	97.6	90.9	67.2	99.0	97.2	71.7	50.2	90.0	93.6	98.4	95.7	58.3	97.1	99.2	99.4	87.0
UniAD [50]	99.7	95.2	86.9	99.8	99.2	93.7	87.5	94.2	99.8	95.8	99.8	98.2	**100**	99.3	98.6	96.5
UniAD(28×28)	99.8	**96.4**	83.2	99.9	98.2	82.1	90.6	93.3	99.8	94.1	99.9	93.6	**100**	99.1	98.2	95.2
UniAD(56×56)	99.5	95.7	82.3	99.8	98.4	67.5	86.7	93.3	99.8	92.3	**100**	**99.7**	**100**	98.0	97.9	94.1
UniAD+VAE	99.7	95.9	86.7	99.8	98.8	93.3	87.8	93.6	99.7	96.0	99.7	98.1	**100**	99.4	98.5	96.5
UniAD+GAN	99.5	94.6	79.4	99.2	99.0	73.8	86.0	89.7	**99.9**	91.3	99.9	98.7	**100**	97.9	97.7	93.8
UniAD+UNet	99.8	94.6	86.7	99.9	99.3	94.8	86.8	94.7	99.8	95.0	99.8	97.5	**100**	99.2	98.4	96.4
Ours	**100**	94.7	**94.0**	**100**	**99.9**	**96.9**	**95.6**	**100**	99.8	**99.7**	98.5	99.6	**100**	**99.9**	98.7	**98.5**

and output I_{out} to the scale $\frac{1}{l}$ and then upsampling to the original resolution,

$$\mathcal{H}_{\text{diff}}^{t,c,l} = U(||D(I_{\text{ori}}, \frac{1}{l}) - D(I_{\text{out}}^{t,c}, \frac{1}{l})||_2, l) \in \mathbb{R}^{H_{\text{img}} \times W_{\text{img}}}, \tag{16}$$

where $U(\cdot)$, $D(\cdot)$ represent the upsampling and downsampling operation, respectively. Due to the low spatial resolution of the feature reconstruction, multi-scale fusion of the base heatmap does not provide extra performance gains, as will be demonstrated in the ablation study. The spatio-temporal heatmap $\mathcal{H}_{\text{diff}}^{\text{ST}}$ is then computed as the per-pixel average of all generated heatmaps with varying timesteps, multiple scales, and different inpainting specifications,

$$\mathcal{H}_{\text{diff}}^{\text{ST}} = \frac{1}{n_t n_c n_l} \sum_{i=1}^{n_t} \sum_{j=1}^{n_c} \sum_{k=1}^{n_l} \mathcal{H}_{\text{diff}}^{t_i,c_j,l_k} \in \mathbb{R}^{H_{\text{img}} \times W_{\text{img}}}. \tag{17}$$

Since the predicted $I_{\text{out}}^{t,c}$ are obtained by sampling the single reverse diffusion process at different timesteps, they contain some noise along with the information we need, we can smooth it to further optimize its representation. The smoothed spatio-temporal heatmap $\mathcal{H}_{\text{diff}}^{\text{SST}}$ is thus post-processed by a mean-filter convolution,

$$\mathcal{H}_{\text{diff}}^{\text{SST}} = \mathcal{H}_{\text{diff}}^{\text{ST}} * f_{m \times m} \in \mathbb{R}^{H_{\text{img}} \times W_{\text{img}}}, \tag{18}$$

where $f_{m \times m}$ is the mean filter of size $(m \times m)$ used for smoothing, $*$ is the convolution operation. In addition, we set a hyper-parameter γ to adjust the weight on heatmaps,

$$\mathcal{H}_{\text{out}} = (1-\gamma)\frac{\mathcal{H}_{\text{base}}}{C_{\text{feat}}} + \gamma \frac{\mathcal{H}_{\text{diff}}^{\text{SST}}}{C_{\text{img}}}. \tag{19}$$

Anomaly detection aims to detect whether an image contains anomalous regions, which can be obtained by taking the maximum value of the averagely pooled $\mathcal{H}_{\text{out}}$.

5. Implementation Details

All images used in our approach are resized to 224×224 pixels. For the base model, we used a ImageNet [10] pre-trained EfficientNet-b4 [42] as a feature extractor. Features from stages 1 to 4 are selected. Then these features are resized to 14×14, and concatenated along channel dimension to form a multi-scale feature map F_{in}. For our diffusion refinement branch, we use the same UNet [33] architecture as backend model from Dhariwal and Nichol [11] and add a model condition as an extra input for noise approximation. The diffusion probabilistic model was set to 1000 diffusion timesteps for training. During inference, we used the DDIM [41] method for deterministic sampling. For the selection of the time period, the initial index of reverse diffusion is mainly considered to ensure that the samples contain low-frequency information; while the number of equally spaced samples to ensure the output heatmaps are accurate and stable. As a result, we empirically used timesteps of $t \in \{250, 200, 150, 100, 50, 0\}$. Besides, we applied inpainting grid size $c \in \{1, 16, 32\}$, scale size $l \in \{1, 2, 4, 8\}$ and $\gamma = 0.9$, $n_s = 2$, $m = 41$. Both the base model and back-end model of diffusion are trained from the scratch by AdamW [17] with weight decay 1×10^{-4}.

6. Experiment

6.1. Datasets and Metrics

Datasets. MVTec-AD [2] includes 15 sub-categories and total 5,354 images, where 3,629 images are train images which are all normal, and 1,725 test images consist of both normal and anomalous images with ground truth mask. Anomaly images are categorized with various kinds of defects. BeanTechAD/BTAD [25] is also industrial anomaly detection dataset with 3 sub-categories and total 2,830 real-world images, among which 1,800 images are for training.

Table 3. **Anomaly localization results with AUROC metric on MVTec-AD**. All methods are evaluated under the unified case. In the unified case, the learned model is applied to detect anomalies for all categories *without* fine-tuning.

Method	Object										Texture					Avg
	Bottle	Cable	Capsule	Hazelnut	MetalNut	Pill	Screw	Toothbrush	Transistor	Zipper	Carpet	Grid	Leather	Tile	Wood	
Student-Teacher [4]	67.9	78.3	85.5	93.7	76.6	80.3	90.8	86.9	68.3	84.2	88.7	64.5	95.4	82.7	83.3	81.8
PatchSVDD [49]	86.7	62.2	83.1	97.4	96.0	96.5	74.3	98.0	78.5	95.1	78.6	70.8	93.5	92.1	80.7	85.6
PaDiM [9]	96.1	81.0	96.9	96.3	84.8	87.7	94.1	95.6	92.3	94.8	97.6	71.0	84.8	80.5	89.1	89.5
FCDD [22]	56.0	64.1	67.6	79.3	57.5	65.9	67.2	60.8	54.2	63.0	68.6	65.8	66.3	59.3	53.3	63.3
MKD [36]	91.8	89.3	88.3	91.2	64.2	69.7	92.1	88.9	71.7	86.1	95.5	82.3	96.7	85.3	80.5	84.9
DRÆM [54]	87.6	71.3	50.5	96.9	62.2	94.4	95.5	97.7	64.5	98.3	98.6	**98.7**	97.3	**98.0**	**96.0**	87.2
AST [35]	76.3	91.3	92.6	92.2	87.7	67.4	82.5	95.5	93.0	79.5	94.8	85.7	88.6	88.1	77.0	86.1
SimpleNet [21]	94.3	91.3	95.3	95.0	91.0	88.9	93.0	95.2	91.0	96.3	97.6	49.5	97.1	93.6	88.8	90.5
UniAD [50]	98.1	**97.3**	98.5	98.1	94.8	95.0	98.3	98.4	97.9	96.8	98.5	96.5	98.8	91.8	93.2	96.8
UniAD(28×28)	97.9	97.0	98.2	98.4	94.4	90.7	98.7	98.5	96.1	96.7	98.5	86.5	99.1	89.3	93.5	95.6
UniAD(56×56)	97.6	96.6	97.9	98.5	94.0	90.0	98.1	98.2	96.8	96.1	**98.7**	94.7	**99.2**	88.9	93.5	95.9
UniAD+VAE	98.0	**97.3**	98.4	98.2	93.5	94.9	98.4	98.4	98.2	96.5	98.5	96.5	98.9	92.0	93.3	96.7
UniAD+GAN	98.2	96.7	97.0	98.1	94.0	90.3	97.4	98.3	98.1	95.2	98.5	94.7	99.0	90.1	93.4	95.9
UniAD+UNet	98.1	**97.3**	98.5	98.0	93.2	95.2	98.3	98.4	**98.4**	96.4	98.5	96.4	98.8	91.9	93.3	96.7
Ours	**98.5**	97.2	**99.0**	**98.8**	**97.5**	**98.3**	**99.5**	**98.9**	97.6	**98.9**	**98.7**	**98.7**	**99.2**	95.0	95.8	**98.1**

Metrics. Following prior works [2, 4, 54], the Area Under the Receiver Operating Curve (AUROC) is used as the evaluation metric for both anomaly detection and localization.

6.2. Comparisons to State-of-the-Art Methods

Baselines. Our approach is compared with baselines including: Student-Teacher [4], PatchSVDD [49], PaDiM [9], MKD [36], DRÆM [54], AST [35], SimpleNet [21]. Note that all of these one-model-per-class methods above are compatible with multi-class input. We've adapted them to manage such input to facilitate a comparative performance analysis. Given the scarcity of methods designed for managing multi-class input, we've also drawn comparisons with UniAD [50] and modified several variants. These include two variants for high-resolution feature reconstruction 28×28, 56×56, and three for raw-resolution image reconstruction (vanilla VAE [18], vanilla GAN [13], UNet [33], respectively). The anomaly heatmaps from these variants are fused with the base heatmap of the feature reconstruction to generate the final anomaly score. All competitors are run with publicly available implementations.

Results of Anomaly Detection on MVTec-AD are shown in Tab. 2. Our proposed framework achieves excellent performances under most categories, and surpasses the strongest competitor by 2.0%, demonstrating our superiority. Especially in categories with more detail defects such as screws or capsules, our DRDC beat the third competitor by a large margin (5.0%∼7.1%). In comparison with various variants of UniAD, we found that simply increasing the resolution of feature vectors does not improve the ability to detect anomalies, but rather weakens the Transformer's ability to correlate with context. As the number of feature tokens grows and the number of training samples remains limited, the model has difficulty finding correlations between features, leading to overfitting as well as deteriorating performance. For the variant that integrates with VAE or GAN, these methods reconstruct the high-resolution image from the latent code. However, the loss of spatial information is inevitable, and training instability arises when combined with GAN. The results indicate that the image anomaly detection performance of UniAD+VAE is not up to par, while the performance of UniAD+GAN even greatly degraded compared to the baseline. The direct combination of UNet is also not effective in improving the performance, since the UNet image reconstruction still needs to trade the image resolution (e.g. pooling and strided convolutions) for semantic information. In contrast, although the back-end model of our method is still UNet, the final performance is improved by 1.4% because the diffusion refinement always operates on the original resolution and the introduced inpainting effectively avoids "identical shortcuts".

Results of Anomaly Localization on MVTec-AD are illustrated in Tab. 3. Although the improvement of the results in anomaly localization is relatively smaller, it still achieves a 2.7% improvement in metal nuts and a 3.3% improvement in pills. Furthermore, compared to those variants with high-resolution feature reconstruction and those methods that incorporate VAE, GAN, and UNet for high-resolution image reconstruction, our approach also exhibits higher performance in anomaly localization, and these slight but stable improvements again reflect the effectiveness of our method. The improvement in anomaly localization performance is mainly due to the high-frequency refinement of image reconstruction performed by our modified diffusion.

Results on BeanTechAD are reported in Tab. 4. The high-resolution variants of UniAD do not achieve an advantage in this dataset and even show performance degradation. In contrast, we achieve 93.59 image-wise AUROC and 98.26 pixel-wise AUROC which still stably outperform the aver-

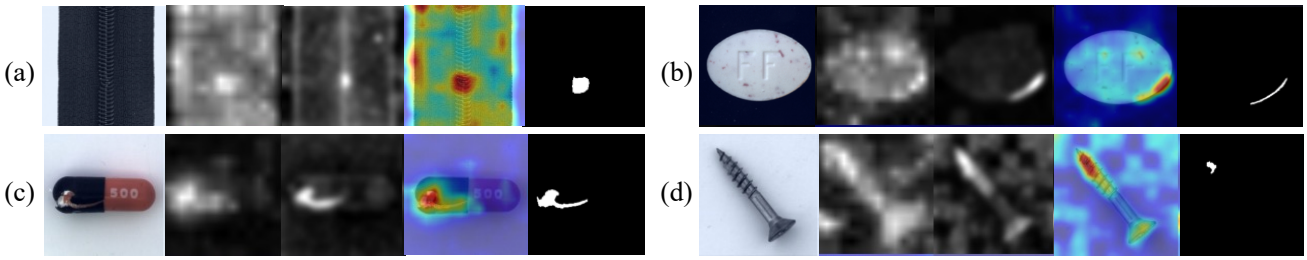

Figure 2. Examples of our proposed framework. Each case from left to right is the original image, the heatmap of the base model, the heatmap of the diffusion refinement, the final synthesized heatmap with the original image, and the ground truth mask, respectively. Note that for illustration purposes, we have only chosen $\mathcal{H}_{\text{diff}}$ at timestep $t=50$ as the example, while the output will fuse all generated heatmaps.

Table 4. **Anomaly detection results with Image-wise AUROC (I-AU) and Pixel-wise AUROC (P-AU) on BTAD** under the unified case.

Method	1		2		3		Avg	
	I-AU	P-AU	I-AU	P-AU	I-AU	P-AU	I-AU	P-AU
UniAD	99.51	97	76.96	95.14	98.83	98.63	91.77	96.92
UniAD(56×56)	99.12	97.55	78.1	95.06	98.84	98.7	92.02	97.1
UniAD+VAE	99.77	96.92	77.33	94.9	99.81	**99.68**	92.3	97.17
UniAD+GAN	98.15	96.95	77.95	94.69	**99.82**	99.57	91.97	97.07
UniAD+UNet	97.95	97.02	78.78	94.81	98.97	99.46	91.9	97.13
Ours	**99.81**	**98.53**	**82.07**	**96.75**	98.88	99.5	**93.59**	**98.26**

Table 5. **Ablation studies with AUROC metric on MVTec-AD** under the unified case.

Multi-scale in base heatmap	×	×	×	✓	×
Spatio-temporal fusion	✓	✓	×	✓	✓
Model condition	×	✓	✓	✓	✓
Test-time condition	✓	×	✓	✓	✓
Image-wise AUROC	97.74	98.00	97.80	97.56	**98.45**
Pixel-wise AUROC	98.02	97.82	97.98	97.69	**98.05**

age performance on UniAD by ≈2.0%. In particular, samples of the 2-category have a relatively lower original performance and our DRDC get a larger boost, even outperforming the one-model–per-category methods [9, 25, 30], indicating the efficacy of our proposed framework.

Visualization on MVTec-AD are shown in Fig. 2. The base model's heatmap is notably blurry, with indistinct boundaries and some activation cases present in the background region. This is primarily due to its low spatial resolution, which hampers the performance. However, our diffusion-based method, which predict noise and operate on the original image, naturally maintain the fineness of the original resolution. With the help of reverse diffusion process, the edges of the generated heatmap are clearer. The final heatmaps, obtained through spatio-temporal fusion, are comparatively pure. This is attributed to two factors: first, the diffusion process operates on the original resolution, resulting in less spatial information loss; second, we employ dual conditions to boost reconstruction accuracy and use inpainting to further avoid "identical shortcuts".

Histogram of the probability distribution of our DRDC method and UniAD [baseline] for each category on the MVTec-AD dataset are shown in the supplementary material, which demonstrates that our DRDC exhibits clearer boundaries between normal and abnormal samples.

6.3. Ablation Studies

To substantiate the effectiveness of our proposed modules, we conducted comprehensive ablation studies on the MVTec-AD dataset. The results, presented in Tab. 5, offer a comparative analysis of the performance of different algorithm variants, each evaluated with the addition and removal of individual modules. Combined with UniAD's original performance, we observe that the mere application of test-time conditions and spatio-temporal fusion can enhance the model's detection and localization by 1.2%. The inclusion of the model condition further boosts the performance improvement to 2.0%. However, integrating a multi-scale algorithm into the base heatmap led to a 1.0% decline in results. Consequently, we opted not to incorporate the multi-scale algorithm into the base heatmap in our final implementation. The ablation study results underscore the effective contribution of each proposed module to the algorithm.

We also performed a sensitivity analysis on hyperparameters and discovered that most of them are insensitive, ensuring the model's high performance remains consistent. Due to space constraints, we have included the results and running time in the supplementary material.

7. Conclusion

We discovered the low-resolution issue in previous anomaly detection reconstruction-based models and proposed an noval framework by exploiting the diffusion model for refinement. The diffusion model is employed for the inpainting task to circumvent the "identical shortcuts" problem. To increase its sampling speed and takes full advantage of the reverse diffusion process, our method reconstructs only the high-frequency part to refine the original heatmap. To maintain accuracy and accommodate multi-class anomaly detection settings, we introduce dual conditions for category-awareness and develop a spatio-temporal fusion for smoother integration. Extensive experiments showcase the effectiveness of our method, and the contributions of each module are also carefully validated.

References

[1] Samet Akcay, Amir Atapour-Abarghouei, and Toby P Breckon. GANomaly: Semi-supervised anomaly detection via adversarial training. In *ACCV*, 2018. 1

[2] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. MVTec AD–A comprehensive real-world dataset for unsupervised anomaly detection. In *CVPR*, 2019. 1, 2, 3, 6, 7

[3] Paul Bergmann, Sindy Löwe, Michael Fauser, David Sattlegger, and Carsten Steger. Improving unsupervised defect segmentation by applying structural similarity to autoencoders. In *International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP)*, 2019. 1, 2

[4] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Uninformed students: Student-teacher anomaly detection with discriminative latent embeddings. In *CVPR*, 2020. 1, 5, 6, 7

[5] Sam Bond-Taylor, Adam Leach, Yang Long, and Chris G. Willcocks. Deep generative modelling: A comparative review of vaes, gans, normalizing flows, energy-based and autoregressive models. *IEEE TPAMI*, 44(11):7327–7347, 2022. 2

[6] Andrew Brock, Theodore Lim, J.M. Ritchie, and Nick Weston. Neural photo editing with introspective adversarial networks. In *International Conference on Learning Representations*, 2017. 1

[7] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *International Conference on Learning Representations*, 2019. 1

[8] Jooyoung Choi, Sungwon Kim, Yonghyun Jeong, Youngjune Gwon, and Sungroh Yoon. Ilvr: Conditioning method for denoising diffusion probabilistic models. in 2021 ieee. In *ICCV*, pages 14347–14356, 2021. 5

[9] Thomas Defard, Aleksandr Setkov, Angelique Loesch, and Romaric Audigier. PaDim: A patch distribution modeling framework for anomaly detection and localization. In *ICPR*, 2021. 1, 2, 6, 7, 8

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, 2009. 6

[11] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *NeurIPS*, 34:8780–8794, 2021. 2, 6

[12] Tharindu Fernando, Harshala Gammulle, Simon Denman, Sridha Sridharan, and Clinton Fookes. Deep learning for medical anomaly detection–a survey. *ACM Computing Surveys (CSUR)*, 2021. 1

[13] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*. Curran Associates, Inc., 2014. 1, 7

[14] S. Gu, D. Chen, J. Bao, F. Wen, B. Zhang, D. Chen, L. Yuan, and B. Guo. Vector quantized diffusion model for text-to-image synthesis. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10686–10696, Los Alamitos, CA, USA, 2022. IEEE Computer Society. 5

[15] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 33:6840–6851, 2020. 1, 3, 4

[16] Chaoqin Huang, Haoyan Guan, Aofan Jiang, Ya Zhang, Michael Spratlin, and Yanfeng Wang. Registration based few-shot anomaly detection. In *ECCV*, 2022. 2

[17] Loshchilov Ilya and Hutter Frank. Decoupled weight decay regularization. In *ICLR*, 2019. 6

[18] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 7

[19] Chun-Liang Li, Kihyuk Sohn, Jinsung Yoon, and Tomas Pfister. CutPaste: Self-supervised learning for anomaly detection and localization. In *CVPR*, 2021. 1, 6

[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 *CVPR*, 2020. 1, 2

[21] Zhikang Liu, Yiming Zhou, Yuansheng Xu, and Zilei Wang. Simplenet: A simple network for image anomaly detection and localization, 2023. 6, 7

[22] Philipp Liznerski, Lukas Ruff, Robert A Vandermeulen, Billy Joe Franks, Marius Kloft, and Klaus-Robert Müller. Explainable deep one-class classification. In *ICLR*, 2021. 1, 2, 7

[23] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11461–11471, 2022. 5

[24] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. In *ICLR*, 2021. 5

[25] 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 *International Symposium on Industrial Electronics*, 2021. 2, 6, 8

[26] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations*, 2018. 1

[27] Arian Mousakhan, Thomas Brox, and Jawad Tayyub. Anomaly detection with conditioned denoising diffusion models. *arXiv preprint arXiv:2305.15956*, 2023. 2

[28] Pramuditha Perera, Ramesh Nallapati, and Bing Xiang. OCGAN: One-class novelty detection using GANs with constrained latent representations. In *CVPR*, 2019. 1

[29] Walter Hugo Lopez Pinaya, Petru-Daniel Tudosiu, Robert Gray, Geraint Rees, Parashkev Nachev, Sébastien Ourselin, and M. Jorge Cardoso. Unsupervised brain anomaly detection and segmentation with transformers. In *Medical Imaging with Deep Learning*, 2021. 1

[30] Tal Reiss, Niv Cohen, Liron Bergman, and Yedid Hoshen. Panda: Adapting pretrained features for anomaly detection and segmentation. In *CVPR*, 2021. 8

[31] Oliver Rippel, Patrick Mertens, and Dorit Merhof. Modeling the distribution of normal data in pretrained deep features for anomaly detection. In *ICPR*, 2021. 2

[32] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021. 1

[33] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. pages 234–241. Springer, 2015. 6, 7

[34] Marco Rudolph, Tom Wehrbein, Bodo Rosenhahn, and Bastian Wandt. Fully convolutional cross-scale-flows for image-based defect detection. 2022. 5

[35] Marco Rudolph, Tom Wehrbein, Bodo Rosenhahn, and Bastian Wandt. Asymmetric student-teacher networks for industrial anomaly detection. 2023. 6, 7

[36] Mohammadreza Salehi, Niousha Sadjadi, Soroosh Baselizadeh, Mohammad H Rohban, and Hamid R Rabiee. Multiresolution knowledge distillation for anomaly detection. In *CVPR*, 2021. 5, 6, 7

[37] Florian Schmidt. Generalization in generation: A closer look at exposure bias. In *Conference on Empirical Methods in Natural Language Processing*, 2019. 5

[38] Bernhard Schölkopf, John C Platt, John Shawe-Taylor, Alex J Smola, and Robert C Williamson. Estimating the support of a high-dimensional distribution. *Neural Computation*, 2001. 2

[39] Yong Shi, Jie Yang, and Zhiquan Qi. Unsupervised anomaly segmentation via deep feature reconstruction. *Neurocomputing*, 2021. 1

[40] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. pages 2256–2265. PMLR, 2015. 3

[41] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv:2010.02502*, 2020. 4, 6

[42] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. pages 6105–6114. PMLR, 2019. 6

[43] Aaron van den Oord, Oriol Vinyals, and koray kavukcuoglu. Neural discrete representation learning. In *NeurIPS*. Curran Associates, Inc., 2017. 1

[44] Julia Wolleb, Florentin Bieder, Robin Sandkühler, and Philippe C Cattin. Diffusion models for medical anomaly detection. In *International Conference on Medical image computing and computer-assisted intervention*, pages 35–45. Springer, 2022. 2

[45] Jhih-Ciang Wu, Ding-Jie Chen, Chiou-Shann Fuh, and Tyng-Luh Liu. Learning unsupervised metaformer for anomaly detection. In *ICCV*, pages 4369–4378, 2021. 2

[46] Julian Wyatt, Adam Leach, Sebastian M. Schmon, and Chris G. Willcocks. Anoddpm: Anomaly detection with denoising diffusion probabilistic models using simplex noise. In *CVPRW*, pages 650–656, 2022. 2

[47] Yingda Xia, Yi Zhang, Fengze Liu, Wei Shen, and Alan L Yuille. Synthesize then compare: Detecting failures and anomalies for semantic segmentation. In *ECCV*, 2020. 1

[48] Zhisheng Xiao, Karsten Kreis, and Arash Vahdat. Tackling the generative learning trilemma with denoising diffusion GANs. In *ICLR*, 2022. 2

[49] Jihun Yi and Sungroh Yoon. Patch SVDD: Patch-level SVDD for anomaly detection and segmentation. In *ACCV*, 2020. 1, 6, 7

[50] Zhiyuan You, Lei Cui, Yujun Shen, Kai Yang, Xin Lu, Yu Zheng, and Xinyi Le. A unified model for multi-class anomaly detection. In *NeurIPS*, 2022. 1, 2, 6, 7

[51] 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. 2

[52] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Free-form image inpainting with gated convolution. In *ICCV*, pages 4471–4480, 2019. 5

[53] Muhammad Zaigham Zaheer, Jin-ha Lee, Marcella Astrid, and Seung-Ik Lee. Old is gold: Redefining the adversarially learned one-class classifier training paradigm. In *CVPR*, 2020. 1

[54] Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj. DRÆM-A discriminatively trained reconstruction embedding for surface anomaly detection. In *ICCV*, 2021. 1, 6, 7

[55] Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj. Reconstruction by inpainting for visual anomaly detection. *PR*, 112:107706, 2021. 5