DriftRemover: Hybrid Energy Optimizations for Anomaly Images Synthesis and Segmentation

Siyue Yao[1,2], **Haotian Xu**[3], **Mingjie Sun**[4*], **Siyue Yu**[1], **Jimin Xiao**[1], **Eng Gee Lim**[1]
[1]Xi'an Jiaotong-Liverpool University
[2]University of Liverpool
[3]RippleInfo
[4]Soochow University

Abstract

This paper tackles the challenge of anomaly image synthesis and segmentation to generate various anomaly images and their segmentation labels to mitigate the issue of data scarcity. Existing approaches employ the precise mask to guide the generation, relying on additional mask generators, leading to increased computational costs and limited anomaly diversity. Although a few works use coarse masks as the guidance to expand diversity, they lack effective generation of labels for synthetic images, thereby reducing their practicality. Therefore, our proposed method simultaneously generates anomaly images and their corresponding masks by utilizing coarse masks and anomaly categories. The framework utilizes attention maps from synthesis process as mask labels and employs two optimization modules to tackle drift challenges, which are mismatches between synthetic results and real situations. Our evaluation demonstrates that our method improves pixel-level AP by 1.3% and F1-MAX by 1.8% in anomaly detection tasks on the MVTec dataset. Additionally, its successful application in practical scenarios highlights its effectiveness, improving IoU by 37.2% and F-measure by 25.1% with the Floor Dirt dataset. The code is available at *https://github.com/JJessicaYao/DriftRemover*.

1 Introduction

Anomaly Images Synthesis and Segmentation (AISS) task aims to generate various types of anomaly images and their segmentation labels for downstream tasks [Li *et al.*, 2021; Lin *et al.*, 2021; Duan *et al.*, 2023; Hu *et al.*, 2023]. The emergence of AISS task is due to the scarcity of real-world anomaly samples, such as industrial inspection and home cleaning scenarios, making it challenging for downstream methods to locate anomalous regions that deviate from the normal areas in images. In this way, AISS can significantly increase the quantity and diversity of anomaly datasets for downstream anomaly localization and segmentation methods.

*Corresponding author (mjsun@suda.edu.cn).

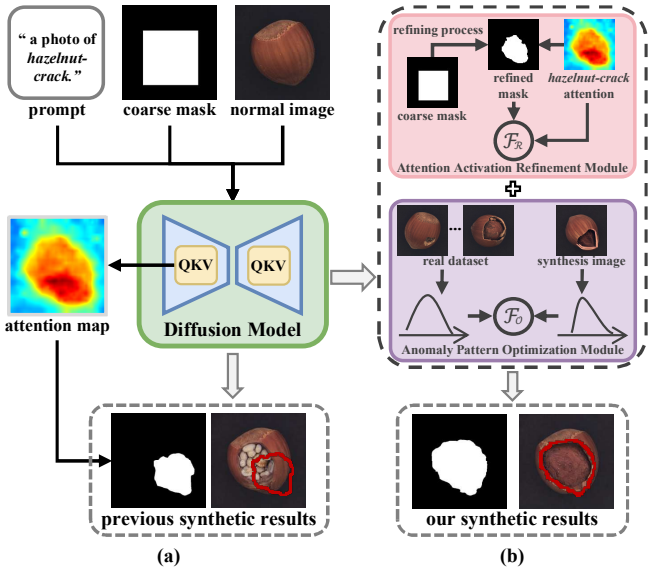

Figure 1: Framework comparison between the baseline method and ours. (a) The baseline method directly uses attention maps to create mask labels of synthetic anomalies but suffers from artificial appearance and mismatches. (b) Two modules are proposed in our method to alleviate these issues. The Attention Activation Refinement module (depicted in purple) compels the attention map to converge with the refined mask, improving consistency between the anomaly region and its label. The Anomaly Pattern Optimization module (depicted in pink) ensures that generated anomaly distribution closely aligns with the real dataset, enhancing realism of generated images.

Previous AISS methods, such as AnomalyDiffusion [Hu *et al.*, 2023], heavily rely on the guidance mask, created by additional models, to generate the image with the anomalous region located within this guidance mask, which increases the computational costs and limits synthetic anomaly diversity. To overcome this limitation, AnoGen [Guan Gui, 2024] employs the coarse mask (*e.g.*, bounding box), rather than the precise mask, to guide generation process, enhancing synthetic diversity while reducing computational demands. However, this method cannot obtain precise mask labels for synthetic anomalous images, limiting the applicability of synthetic samples in downstream tasks. To solve this issue, we

propose a new pipeline to extract cross-attention maps during anomalies generation, used as masks after binarization, as shown in Figure 1(a). These paired samples serve as images and labels for training in downstream tasks to enhance performance. Nevertheless, simply employing this naive pipeline as a synthetic baseline poses significant drift challenges.

The first drift challenge is the misalignment between generated mask label and synthetic anomaly region. As illustrated in Figure 1(a), the black-and-white image in the lower left corner clearly shows a discrepancy between the mask derived from the attention layer and the generated image in the lower right corner. The reason for the discrepancy is that when generating anomalies with keywords like "crack", the attention map activates broad regions and is accompanied by vague boundaries, leading to inaccurately binarized masks.

Another drift challenge is the distribution discrepancies between generated images and real-world scene images. This issue arises from the diffusion model's tendency to misunderstand ambiguous terms, leading to significant pattern and style inconsistency between synthetic and real anomalies [Kim *et al.*, 2024]. For instance, in Figure 1(a), the synthetic anomalies are represented as a cluster of small white circles, however, the intended anomaly should depict a broken hazelnut, which fails to accurately represent the nature of real anomalies. Consequently, this drift limits the utility of synthetic data for training purposes in downstream tasks.

To address the aforementioned challenges, we undertake additional optimization of the proposed baseline by introducing two novel modules in the synthesis process. These modules specifically solve drift challenges of label mismatches and pattern discrepancies, which are critical for enhancing the authenticity of anomalies and the accuracy of masks.

The first module, Attention Activation Refinement, is meticulously designed to guarantee the congruence between the synthesised anomaly regions and their corresponding segmentation masks, as shown in the pink part of Figure 1(b). It initiates with a preliminary coarse mask, which is then subjected to an advanced refining process to generate a refined mask. By increasing the attention map's activation values within the refined mask region, the attention map achieves more distinct edges, facilitating the production of binarized masks that are more precisely aligned with the anomaly region. This refinement leads to a significant improvement in the accuracy of the mask's shape and position, surpassing the initial coarse mask's accuracy. The accurate mask synthesised by our method in Figure 1(b) corroborates this result.

The second module, Anomaly Pattern Optimization, aligns the category distribution of generated images with anomaly regions in real-world scenarios. Specifically, the method gathers the distribution of anomaly regions for various categories in the existing dataset. It minimizes the difference between the real and generated anomaly patterns to improve the realism of the synthesised anomaly images by making synthetic anomaly region distributions match the real corresponding category distributions. For instance, in the purple part of Figure 1(b), the image synthesised by our method can better reproduce the internal appearance of hazelnuts.

By harmoniously integrating these two optimization modules with the previously proposed baseline, our method, DriftRemover, effectively addresses AISS task and provides high-quality training data, which is crucial for subsequent anomaly localization and segmentation tasks. Our contributions can be encapsulated within the following three aspects:

- Our DriftRemover synthesises anomaly images conditioned on coarse masks and anomaly categories, while generates precise labels utilizing an attention mechanism. This effectively synthesises diverse anomaly images while maintaining high accuracy in labelling, thereby enhancing the downstream task performance.
- Two optimization modules is introduced to resolve drift issues in synthesis process: Attention Activation Refinement (AAR) and Anomaly Pattern Optimization (APO). AAR dynamically refines attention maps for precise segmentation masks, while APO enhances realism of synthesised anomalies by aligning their style with real ones.
- Extensive experiments demonstrate that our generated data effectively improves the performance of downstream tasks. On the MVTec dataset, AP and F1-MAX increase by 1.3% and 1.8%. On the Floor Dirt dataset, IoU and F-measure increase 37.2% and 25.1%.

2 Related Work

2.1 Anomaly Detection

Anomaly detection aims to identify samples that significantly deviate from the normal distribution, indicating anomalies. Previous methods fall into three categories: Reconstruction-based methods [Gong *et al.*, 2019; Park *et al.*, 2020; Ristea *et al.*, 2022; Schlüter *et al.*, 2022] analyse the residuals between input and reconstructed output, relying on the model's ability to reconstruct normal regions while struggling with abnormal ones. Embedding-based methods [Roth *et al.*, 2022; Yao *et al.*, 2023; Zhang *et al.*, 2023c; Liu *et al.*, 2023], including memory banks, distribution map, teacher-student and one-class classification, utilize pre-trained networks to extract features and compress normal features to separate anomalies, but they lack direct learning of anomaly features. Augmentation-based methods use synthetic anomaly images for training. Image-level approaches [Lin *et al.*, 2021; Zavrtanik *et al.*, 2021] provide detailed anomaly textures but lack diversity, while feature-level methods [Yan *et al.*, 2021; You *et al.*, 2022] are efficient but face challenges in control.

2.2 Anomaly Synthesis

The limited availability of anomaly data sparks significant interest in anomaly generation. Previous methods [Zavrtanik *et al.*, 2021; Zhang *et al.*, 2023a] use Perlin noise and cut-paste techniques but lack realism and diversity. Inspired by GANs [Goodfellow *et al.*, 2014], researchers explore methods [Niu *et al.*, 2020; Zhang *et al.*, 2021] that require defect-free samples as input. DFMGAN [Duan *et al.*, 2023] generates anomaly images and masks but struggles with complex objects. While diffusion models are widely used for image generation [Ho *et al.*, 2020; Song and Ermon, 2019; Zhang *et al.*, 2023b], recent approaches [Hu *et al.*, 2023; Qiu *et al.*, 2025] require separate models for precise masks, resulting in high computational costs. Moreover, AnoGen

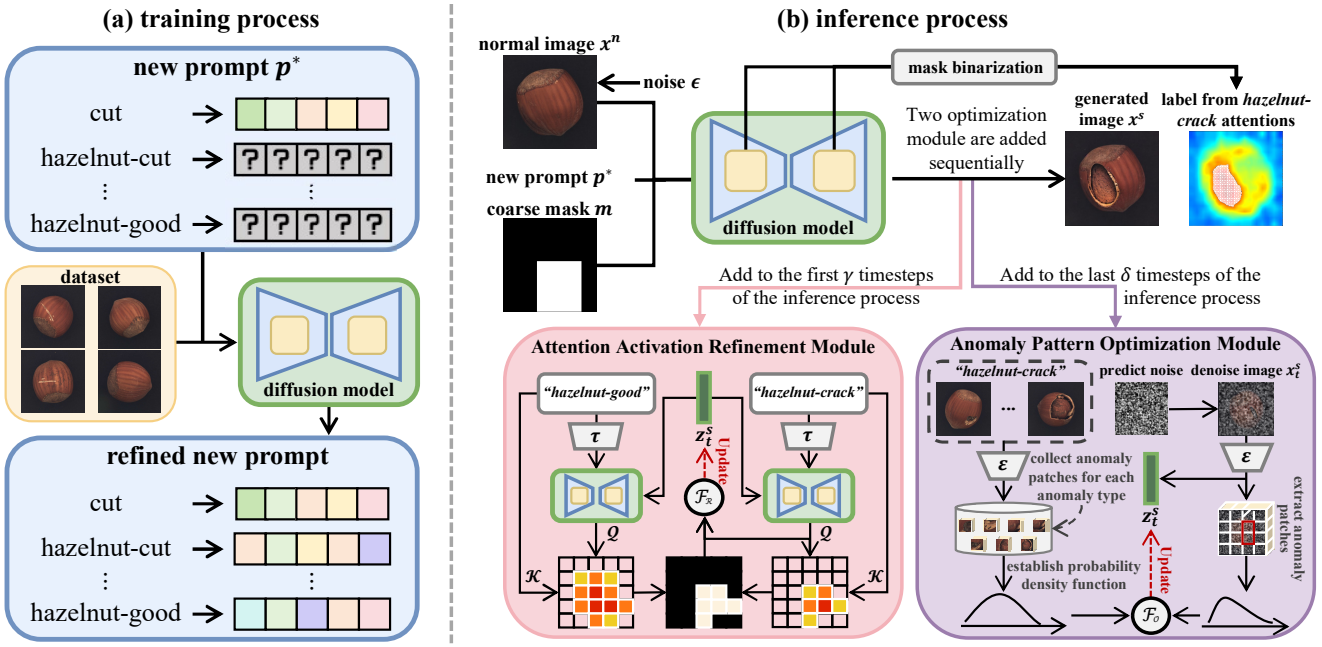

Figure 2: Illustration of the framework. (a) During training, the proposed method learns the textual embedding corresponding to each anomaly category by optimizing the generated results of the diffusion model. (b) Early in inference, the Attention Activation Refinement module amplifies the activation values within the attention map of the anomaly region, ensuring that the generated mask labels closely align with the actual anomalies. Later, the Anomaly Pattern Optimization module is further proposed, which applies a data distribution model based on real anomaly datasets to refine the appearances and styles of synthetic anomaly images, enhancing their realism.

[Guan Gui, 2024] uses bounding boxes to eliminate the need for a mask generator but fails to synthesize precise masks.

3 Method

The proposed DriftRemover pipeline, as depicted in Figure 2, effectively synthesises anomaly images while generating mask labels. During training, it employs learnable embeddings to extract features from various anomalies, aiding in both image synthesis and mask extraction. In inference, the Attention Activation Refinement module adjusts attention values to better align with anomaly regions, while the Anomaly Pattern Optimization module enhances the realism of the synthetic anomalies. This approach produces high-quality synthetic data and accurately matched mask labels.

3.1 Preliminary

The AISS task recently adopts the state-of-the-art (SOTA) generative model Stable Diffusion (SD) [Rombach *et al.*, 2022]. During the training stage, SD first encodes the real anomaly image x^r into the latent space via an image encoder $\varepsilon(\cdot)$, resulting in a latent vector $z_0^r = \varepsilon(x^r)$ with the shape of $W \times H \times C$, where W and H represent the size, and C is the number of channels. Afterward, Gaussian noise is gradually added to z_0^r in the forward process, as follows:

$$z_t^r = \sqrt{\bar{\alpha}_t} z_0^r + \sqrt{1-\bar{\alpha}_t} I, \tag{1}$$

where z_t^r is the noisy latent vector, and α_t is the standard deviation of the noise at timestep t. $\bar{\alpha}_t$ represents the cumulative value from α_0 to α_t. I denotes the identity matrix. Finally, a U-Net model is used to estimate the noise increment at any step, and the objective is to minimize the difference between the output of the U-Net model and the added noise.

During inference, SD converts a normal image x^n into a latent vector z_0^n, and then performs a T step forward diffusion process akin to Equation 1 to obtain z_T^n. In reverse, the well-trained U-Net model estimates the noise increment at each step, progressively inferring the latent vector with anomaly features z^s from the noise z_T^n. Simultaneously, a text encoder converts input prompts into embeddings $p \in \mathbb{R}^{l \times q}$ with l tokens, which serve as conditions to guide U-Net, as follows:

$$z_{t-1}^s = \frac{1}{\sqrt{\alpha_t}} \left(z_t^s - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \epsilon_\theta \left(z_t^s, p, t \right) \right) + \sqrt{\beta_t} I, \tag{2}$$

where q is the dimension of text feature; $\epsilon_\theta(z_t^s, p, t)$ is the predicted noise at timestep t; β_t is defined as $1 - \alpha_t$.

The cross-attention layer in SD takes three inputs: a query vector $\mathbf{Q}$, a key vector $\mathbf{K}$, and a value vector $\mathbf{V}$. $\mathbf{K}$ and $\mathbf{V}$ are computed from linear projections of the textual embeddings p, and $\mathbf{Q}$ is sourced from a linear projection of each convolution block output. By combining the vectors $\mathbf{Q}$ and $\mathbf{K}$ in different layers, the cross-attention map $\mathcal{A}$ is as follows:

$$\mathcal{A} = \sum_k \text{softmax}\left(\frac{\mathbf{Q}_k \mathbf{K}_k^\top}{\sqrt{d}}\right), \tag{3}$$

where $\sqrt{d}$ is a scaling factor, usually set to 1. For each layer, the map shape is $\frac{W}{k} \times \frac{H}{k} \times h \times l$, where k is the reduction factor associated with the block's resolution and h is head number. These matrices can be used to weight the image pixels, thereby influencing the final generated anomaly images.

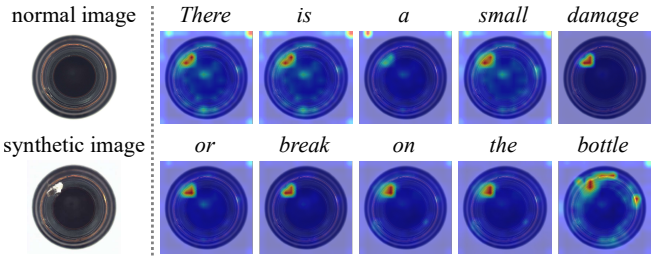

Figure 3: Stable diffusion's cross-attention maps for different single words of the input prompt in the last inference step.

3.2 Mask Labels Synthesis via Attention Module

Cross-attention map reflects the model's focus on specific areas of an image based on the input text, allowing it to locate and distinguish objects. Typically, object category labels are used as text to activate cross-attention maps and then create masks [Marcos-Manchón *et al.*, 2024; Nguyen *et al.*, 2024; Zhao *et al.*, 2023]. However, most anomaly categories include multiple adjectives or verbs instead of simple nouns, making it difficult for activation maps to accurately focus on specific regions. To explore these words effects, we reorganize anomaly categories into descriptive statements to generate different activation maps. For example, the anomaly category "broken_small" expands to "There is a small damage or break on the bottle". Figure 3 reveals a key issue: **no single word effectively delineates the triggering factors of anomalies**. Therefore, there is an urgent need to introduce new keywords or embeddings to accurately capture anomaly category features, which is used to activate attention maps.

Therefore, learnable embeddings are used to capture the semantic features of different types for various objects, such as "hazelnut_good" and "bottle_broken_small". These embeddings are continuously updated until the SD model can effectively generate the corresponding distribution of anomalies. Specifically, an embedding p^* for any anomaly type is established to replace the text conditions of the SD model, which can be optimized similarly to the LDM loss [Rombach *et al.*, 2022]. Besides, to ensure that the learned text embeddings focus exclusively on the region defined by the coarse mask m without influence from external areas, the model input is constructed as the concatenation of the noisy embedding of the real anomaly image z_t^r, the coarse mask m used to determine the anomaly location, and the embedding of normal region in the anomaly image with coarse mask area removed $\varepsilon(x^r(1-m))$. For simple notation, we still mark this input as z_t^r. Overall, the loss for training embeddings p^* is as follows:

$$p^* = \arg\min_{p \in \mathbb{R}^{l \times q}} \mathbb{E}_{z_t^r, p, t, \epsilon}\left[\left|\left|\epsilon - \epsilon_\theta\left(z_t^r, p, t\right)\right|\right|^2\right], \tag{4}$$

where p represents the text embedding of the anomaly category noun with the size $\mathbb{R}^{l \times q}$ for initializing the learnable p^* to reduce training duration. ϵ represents the noise added to the image. In this way, the text embeddings p^* accurately reflect the characteristic of each anomaly category, thus better activating the corresponding visual representation regions.

Ultimately, the learned embeddings of different anomaly types are used to generate attention maps $\mathcal{A}$ in the final timestep. These maps are aggregated and binarized into a mask using Otsu's method [Otsu, 1979] to automatically find the optimal threshold. This process converts the attention maps into binary anomaly masks, resulting in mask labels paired with synthesised images for use in downstream tasks.

3.3 Dynamic Inference for Anomaly Synthesis

To craft a superior synthetic dataset, relying on attention maps for masks is insufficient due to drift issues, resulting in unrealism and label inconsistency. To alleviate such issues, we propose two modules based on energy function mechanism [Chen *et al.*, 2024]. Specifically, the gradient of energy function $\mathcal{F}(c, z_t^s)$ is computed to update latent variable z_t^s under conditions c, thereby improving generation quality. Therefore, the conditional sampling formula is written as follows:

$$z_t^s \leftarrow z_t^s - \sigma_t \nabla_{z_t^s} \mathcal{F}(c, z_t^s), \tag{5}$$

where $\sigma_t = (1-\alpha_t)/\alpha_t$ is a scale factor. By updating the latent variable, all generated images are influenced by backward guidance. Additionally, our method strategically updates the noise vectors to synthesise anomalies without altering the regions outside the coarse masks, similar to the approach outlined in [Avrahami *et al.*, 2022], as follows:

$$z_{t-1}^s \leftarrow z_{t-1}^s m + (\sqrt{\bar{\alpha}_{t-1}} z_0^n + \sqrt{1-\bar{\alpha}_{t-1}} I)(1-m), \tag{6}$$

where z_{t-1}^s is the latent vector synthesised by Equation 2. $\sqrt{\bar{\alpha}_{t-1}} z_0^n + \sqrt{1-\bar{\alpha}_{t-1}} I$ is equivalent to the latent vector z_t^n, i.e., adding t step noise to the embedding of a normal image.

Attention Activation Refinement Module

The Attention Activation Refinement (AAR) module improves SD's cross-attention maps to pinpoint anomaly regions as shown in the pink area in Figure 2, using new embeddings for "hazelnut-good" and "hazelnut-crack" to activate the entire object region $\mathcal{A}_t^g$ and the rough region of the anomaly $\mathcal{A}_t^s$. The overlap of these two activation maps roughly estimates the anomaly location. To keep synthesised anomalies within initial coarse mask, a refined mask $\hat{m}$ is created by excluding out-of-bounds intersections as follows:

$$\hat{m} = (\mathcal{A}_t^g > \eta) \odot (\mathcal{A}_t^s > \eta) \odot m, \tag{7}$$

where $\odot$ represents element-wise multiplication and η denotes a threshold for binarizing the attention maps to masks.

This mask guides the inference process to produce anomalies that closely match the expected areas. An energy function is tailored to maximize activation within the refined mask and minimize its surroundings on the activation map $\mathcal{A}^s$ of the new anomaly embedding p^*, as follows:

$$\mathcal{F}_R(\mathcal{A}_t^s, \hat{m}) = \frac{\sum_{i \in N} w_{t,i} \mathcal{A}_{t,i}^s - \sum_{j \in \widetilde{N}} \widetilde{w}_{t,j} \mathcal{A}_{t,j}^s}{\sum_{i \in N} w_{t,i} + \sum_{j \in \widetilde{N}} \widetilde{w}_{t,j}}, \tag{8}$$

where t illustrates the timestep. N and $\widetilde{N}$ represent the pixel point set with value 0 and value 1 in $\hat{m}$, representing the normal and anomaly regions. w and $\widetilde{w}$ can be regarded as weight measures and are calculated as follows:

$$w_{t,i} = \left[\frac{\mathcal{A}_{t,i}^s}{\mu}\right], \widetilde{w}_{t,j} = \left[\frac{\mathcal{A}_{t,j}^s}{\widetilde{\mu}}\right], \tag{9}$$

where $\widetilde{\mu}$ and μ represent the average activation values of the anomaly region and the remaining region of the activation map. All activation maps $\mathcal{A}$ are normalized in $[0, 1]$.

Algorithm 1 Inference process of proposed DriftRemover

Input: normal image x^n, coarse mask m, input anomaly prompt p^*, total inference timestep T
Parameter: conditional noise predictor $\epsilon_\theta(\cdot)$, image encoder $\varepsilon(\cdot)$, binary mask threshold η, last timestep γ for adding AAR module, first timestep δ for adding APO module, function repeat times Γ, energy value threshold Θ, pre-defined parameters $\bar{\alpha}_T$, I, β_t and σ_t
Output: the synthetic anomaly image latent z_0^s

$z_0^n = \varepsilon(x^n)$; $\quad z_T^n \sim \mathcal{N}(0, I)$
$z_T^s = \textbf{concat}\big(z_T^n, m, \varepsilon\big(x^n(1-m)\big)\big)$
for $t = T$ to 1 **do**
  $i = 0$
  if $t > \gamma$ **then**
    Obtain the attention map with new normal embedding $\mathcal{A}_t^g$ and new anomaly embedding $\mathcal{A}_t^s$.
    $\hat{m} = (\mathcal{A}_t^g > \eta) \odot (\mathcal{A}_t^s > \eta) \odot m$ ▷ **Equation 7**
    while $i < \Gamma$ and $\mathcal{F}_R(\mathcal{A}_t^s, \hat{m}) < \Theta$ **do**
      $z_t^s \leftarrow z_t^s - \sigma_t \nabla_{z_t^s} \mathcal{F}_R(\mathcal{A}_t^s, \hat{m})$ ▷ **Equation 8**
      $i += 1$
    end while
  else if $t < \delta$ **then**
    while $i < \Gamma$ and $\mathcal{F}_O\big(f(z^r), f(z_t^s)\big) < \Theta$ **do**
      $z_t^s \leftarrow z_t^s - \sigma_t \nabla_{z_t^s} \mathcal{F}_O\big(f(z^r), f(z_t^s)\big)$ ▷ **Equation 10**
      $i += 1$
    end while
  end if
  $z_{t-1}^s = \frac{1}{\sqrt{\alpha_t}}\left(z_t^s - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(z_t^s, p^*, t)\right) + \sqrt{\beta_t} I$ ▷ **Equation 2**
  $z_{t-1}^s \leftarrow z_{t-1}^s m + (\sqrt{\bar{\alpha}_{t-1}} z_0^n + \sqrt{1-\bar{\alpha}_{t-1}} I)(1-m)$ ▷ **Equation 6**
end for
return z_0^s

Anomaly Pattern Optimization Module

The Anomaly Pattern Optimization (APO) module aims to enhance the realism of synthetic anomaly images by minimizing the discrepancy between them and actual anomalies. The process begins with encoding real anomaly images x^r through the image encoder $\epsilon(\cdot)$, yielding features z^r. These features are then segmented into $v \times v$ patches, and their mean and variance are calculated to establish the probability density function $f(z^r)$ of the real anomaly distribution. For synthetic anomaly images, patches are extracted directly from the predicted noise features z_t^s, following the same procedure to determine their probability density. The final step involves computing the KL divergence D_{kl} between the real and synthetic anomaly distributions to measure their similarity:

$$\mathcal{F}_O\Big(f(z^r), f(z_t^s)\Big) = D_{kl}\Big(f(z^r) || f(z_t^s)\Big). \tag{10}$$

By employing the KL Divergence, APO bridges the gap between synthetic and real anomaly images, thereby enhancing the realism of the generated anomaly images.

Efficiency Sampling with Sequential Modules

Using two modules simultaneously updating z_t^s during sampling is inefficient. The generation process is divided into an early "structure generation" phase and a later "detail refinement" phase [Hertz *et al.*, 2022; Mokady *et al.*, 2023]. Thus, we recommend applying the AAR module initially for better anomaly positioning, then switching to the APO module to enhance realism. A repeat strategy is also implemented to strengthen the impact of energy functions. If the energy value dips below a threshold Θ or the iteration cap Γ is hit at timestep t, we cease repetition and finalize z_t^s. Details of upgraded sampling algorithm are shown in Algorithm 1. Moreover, at the final timestep, we execute mask labels synthesis procedures in Section 3.2, using attention maps of learned embeddings to generate labels for the synthesised image.

4 Experiment

4.1 Dataset and Metrics

We evaluate our DriftRemover on MVTec [Bergmann *et al.*, 2019] and Floor Dirt dataset. MVTec's original training set consists of 3,629 normal images without any anomaly, while its original test set contains 467 normal images and 1,258 anomaly images along with their corresponding mask labels for the anomaly areas. Subsequently, followed by [Hu *et al.*, 2023], we randomly select 1/3 of the abnormal images for training DriftRemover and the remaining images are used to test the results of the downstream tasks. The Floor Dirt dataset is collected from robotic vacuum cleaners, containing two types of anomalies: stains on the floor (500 images) and pet faeces on the floor (458 images). In our experiments, 3/5 of anomalous images are randomly selected for training our DriftRemover, and 2/5 are used for testing downstream tasks.

The evaluation metrics are divided into two main categories. First, we assess the diversity and quality of the synthesised anomaly images using IC-LPIPS for diversity and Inception Score (IS), where higher values indicate better results. Kernel Inception Distance (KID) is also used for quality, where lower values indicate better results. Second, we measure the accuracy improvement of downstream methods using metrics like Area Under the Receiver Operating Characteristic Curve (AUROC), Intersection over Union (IoU), Average Precision (AP), and maximum F1 score (F1-MAX), all of which show better performance with higher values.

4.2 Implementation Details

Our pipeline is built on Stable Diffusion V1.5 [Rombach *et al.*, 2022], training it for 2,000 epochs with batch size of 4 and image size of 512. The optimizer AdamW utilizes a scaled learning rate initialized to 1e-4. We use 20 steps and a guidance scale of 3.5 for image generation, producing 1,000 images per class for evaluation and training. The threshold Θ and iteration cap Γ are 0.01 and 5. The last timestep γ for adding AAR module is 600, while the first timestep δ for adding APO module is 300. The binary threshold η is 180, patch size v is 3, text dimension q is 768, head number h is 8 and reduction factors k for each layer are 1, 2 and 4.

Dataset	Method	IS ↑	IC-LPIPS ↑	KID ↓
MVTec	DiffAug	1.58	0.09	0.06
	CDC	1.65	0.07	0.11
	Crop-Paste	1.51	0.14	-
	SDGAN	1.71	0.13	0.28
	Defect-GAN	1.69	0.15	0.11
	DFMGAN	1.72	0.20	0.08
	AnomalyDiffusion	1.80	0.32	0.13
	Ours	**1.83**	**0.32**	**0.06**
Floor Dirt	DFMGAN	3.35	0.22	0.19
	AnomalyDiffusion	3.56	0.23	0.10
	Ours	**3.78**	**0.26**	**0.07**

Table 1: Comparison of generated results on the MVTec and Floor Dirt datasets, tested on 1,000 randomly selected synthetic images.

Method	AP ↑	F1-MAX ↑
DREAM	97.0 / 54.1	94.4 / 53.1
DFMGAN	94.8 / 62.7	94.7 / 62.1
AnomalyDiffusion	**99.7** / 81.4	98.7 / 76.3
Ours	**99.7 / 82.7**	**98.9 / 78.1**

Table 2: Comparison of image-level / pixel-level performance of anomaly detection on MVTec by training a U-Net model on synthetic data from different methods. **Bold** denotes the optimal results.

4.3 Quality of Generated Anomalies and Labels

Quantitatively compared with the methods in Table 1, our approach obtains the IS of 1.83 on MVTec, exceeding the previous SOTA method AnomalyDiffusion [Hu *et al.*, 2023], whose IS score is 1.80. Additionally, our approach significantly improves the KID value, reducing it to 0.06. Although the IC-LPIPS remains at 0.32, our DriftRemover preserves the diversity of the generated results while achieving a high degree of fidelity consistent with real images, thus bolstering the robustness of the downstream tasks. This consistent performance is further validated on the Floor Dirt dataset, underscoring the method's robustness and effectiveness.

Figure 4 presents a qualitative analysis of our method on MVTec and Floor Dirt datasets, comparing it with DFMGAN and AnomalyDiffusion. Our DriftRemover excels at generating images that closely align with the target anomaly pattern, such as accurately depicting a "leather_cut" anomaly as a tear with white debris, unlike the coloured patches from other methods. It effectively addresses drift issues, captures the features of anomalies, and enables the shape and pattern of the anomalous areas to more accurately reflect reality.

Additionally, to validate the accuracy of our synthetic mask labels, we manually annotate the anomalies in the generated images and compare them with our generated labels using IoU metric, as depicted in Figure 5. Our DriftRemover achieves an average IoU of 63.96% across all categories, significantly outperforming the IoU results of previous methods.

4.4 Effectiveness for Downstream Task

Anomaly Detection and Localization Task

On the MVTec dataset, Table 2 benchmarks our method against other generation methods to expand training data for

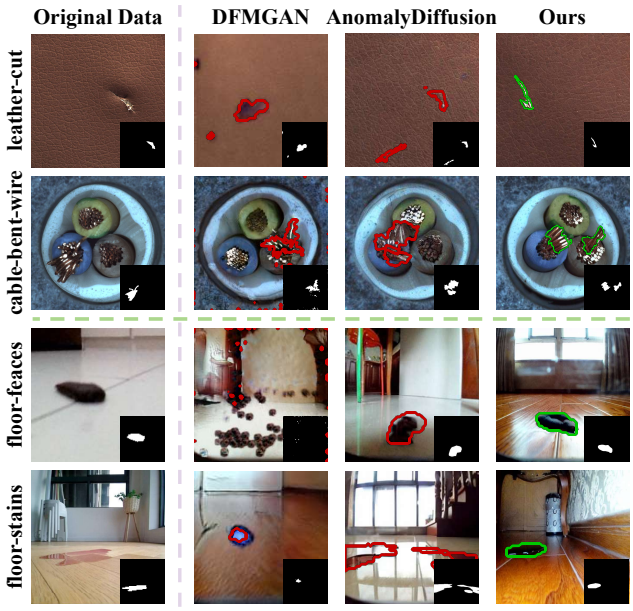

Figure 4: Qualitative comparison on MVTec and Floor Dirt datasets. Mask labels are placed at the bottom right corner.

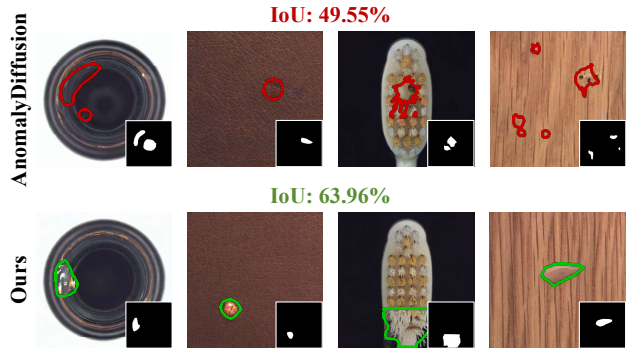

Figure 5: Accuracy comparison of synthetic images and masks. The red line indicates the synthetic mask boundaries, while the manually labelled masks are in the bottom right of each image. The average IoU values between labelled and synthetic masks are shown above.

downstream detection and localization tasks. We compare with Crop-Paste, DRAEM, PRN, DFMGAN, AnomalyDiffusion methods, and synthesise 1,000 images per class to train a simple U-Net for downstream tasks. Our approach achieves 99.7% AP and 98.9% F1-MAX at the image level, and 82.7% pixel-level AP and 78.1% F1-MAX, surpassing competitors by 1.3% and 1.8%, respectively. This indicates the high quality of our synthetic anomalies and masks, enabling simple U-Net to achieve exceptional performance of downstream tasks. Similar AUROC results are observed when compared to both supervised and unsupervised downstream models in Table 3.

Anomaly Segmentation Task

Table 4 evaluates the performance of our DriftRemover against DFMGAN and AnomalyDiffusion on Floor Dirt dataset for anomaly segmentation. We train a SegFormer [Xie *et al.*, 2021] on 1,000 images from each method. Our

Method		Image-level AUROC ↑	Pixel-level AUROC ↑
Unsupervised	CFLOW	97.5	97.7
	DREAM	97.6	96.7
	SSPCAB	97.1	96.3
	CFA	99.1	98.0
	RD4AD	98.7	97.8
	PatchCore	99.2	98.1
Supervised	DevNet	92.2	85.3
	DRA	96.1	85.3
	PRN	99.4	99.0
	Ours	**99.5**	**99.1**

Table 3: Anomaly detection and localization AUROC performance comparison between the model trained on the data generated by our method and previous methods on MVTec dataset.

Method	Category	IoU ↑	F-measure ↑
DFMGAN	stains	39.24	39.00
	feces	37.10	3.45
	average	38.17	21.23
AnomalyDiffusion	stains	35.23	16.00
	feces	34.51	7.07
	average	34.87	11.54
Ours	stains	70.06	38.00
	feces	80.74	54.74
	average	**75.40**	**46.37**

Table 4: Comparison of anomaly segmentation performance on the Floor Dirt dataset, by training a SegFormer model on the data generated by our method and previous methods.

method achieves the highest IoU of 75.40% and F-measure of 46.37%, surpassing other methods and demonstrating superior real-world capabilities in aiding anomaly segmentation.

4.5 Ablation Study

Effect of Different Components

Our main components: learnable anomaly embedding, Attention Activation Refinement (AAR) and Anomaly Pattern Optimization (APO) are evaluated on MVTec, including five settings in Table 5: 1) with none of these components; 2) only learnable anomaly embedding; 3) learnable anomaly embedding and AAR; 4) learnable anomaly embedding and APO; and 5) the full model (ours). For each setting, 1,000 image-mask pairs are first generated for quality evaluation and then used to train the downstream model for the localization evaluation. For setting 1), we directly use the embedding of original keywords (*e.g.*, "crack") to synthesise images and labels while the other settings all use the newly learned embeddings.

The second line of Table 5 shows that the model using the new embeddings can significantly improve the localization task performance. For instance, the pixel-level AP and F1-MAX reach 35.35% and 38.41%, compared to 9.64% and 11.43% for the model using original embedding. This strongly demonstrates the importance and necessity of obtaining embeddings that are tailored to anomaly types, providing a solid data foundation for our future design. Additionally, setting 3) is suitable for the task requiring anomaly localization (83.6% in AP), while setting 4) is suited for applications prioritizing the realism of synthetic images (1.88% in IS). Our model utilizes a balanced setting 5) approach, which achieves a compromise between the generation fidelity (1.83% in IS) and the downstream task performance (82.69% in AP).

	Module ①	②	③	IS ↑	IC-LPIPS ↑	KID ↓	AP ↑	F1-MAX ↑
1)				1.66	0.24	0.07	9.64	11.43
2)	✔			1.62	0.27	0.06	35.35	38.41
3)	✔	✔		1.76	0.34	0.07	83.36	78.87
4)	✔		✔	1.88	0.21	0.05	51.96	51.54
5)	✔	✔	✔	1.83	0.32	0.06	82.69	78.14

Table 5: Main components' contributions on MVTec. ①, ② and ③ denote learnable anomaly embeddings, AAR module and APO module. Only pixel-level metrics are compared for downstream tasks.

Block Resolution			MVTec		Floor Dirt	
64x64	32x32	16x16	AP	F1-MAX	IoU	F-measure
✔			76.00	74.11	67.92	31.22
	✔		75.52	73.50	70.63	39.34
		✔	78.57	76.62	75.25	43.76
✔	✔		78.67	75.70	73.35	43.59
✔		✔	75.71	74.17	**75.94**	44.59
	✔	✔	**82.69**	**78.14**	75.40	**46.37**
✔	✔	✔	77.55	75.13	75.25	43.76

Table 6: Evaluation of different cross-attention resolutions. A U-Net trained on synthetic MVTec images and a SegFormer trained on synthetic Floor Dirt images are evaluated using pixel-level metrics.

Effect of Different Feature Scales

The U-Net model in SD incorporates attention blocks at varying resolutions: 64×64, 32×32, and 16×16, corresponding to reduction factors k of 1, 2 and 4. In our DriftRemover, the 32×32 and 16×16 attention blocks are combined to form the basis of the synthetic mask label. To assess the effect of these attention blocks, we conducted evaluations with various block combinations, as detailed in Table 6. The findings indicate that the 32×32 and 16×16 blocks offer optimal performance. The 16×16 blocks, being in the deeper layers of the U-Net, are adept at processing semantic details, capturing finer-edge information. Conversely, the 32×32 blocks, which operate at a more structural level, provide precise location data of the anomaly regions. The synergistic use of these two resolutions yields an accurate segmentation mask.

5 Conclusion

We present DriftRemover, leveraging coarse masks to synthesise anomaly images and corresponding masks. It utilizes precise anomaly embeddings for attention-guided mask creation and incorporates two optimization modules: Attention Activation Refinement for accurate labels and Anomaly Pattern Optimization for realism. Extensive experiments validate our model's superiority over state-of-the-art methods in realism and its enhancement of downstream task performance. In future, we aim to explore co-optimization of anomaly detection algorithms with synthetic data generation to enhance the feedback loop for creating more effective training samples.

Acknowledgements

This work was supported by Young Scientists Fund of the National Natural Science Foundation of China (Grant No. 62302328), Jiangsu Province Foundation for Young Scientists (Grant No. BK20230482), Suzhou Key Laboratory Open Project (Grant No. 25SZZD07) and Jiangsu Manufacturing Strong Province Construction Special Fund Project (Grant Name: Research and Development and Industrialization of Intelligent Service Robots Integrating Large Model and Multimodal Technology).

References

[Avrahami *et al.*, 2022] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022.

[Bergmann *et al.*, 2019] 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*, 2019.

[Chen *et al.*, 2024] Minghao Chen, Iro Laina, and Andrea Vedaldi. Training-free layout control with cross-attention guidance. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024.

[Duan *et al.*, 2023] Yuxuan Duan, Yan Hong, Li Niu, and Liqing Zhang. Few-shot defect image generation via defect-aware feature manipulation. In *Proceedings of the AAAI conference on artificial intelligence*, 2023.

[Gong *et al.*, 2019] Dong Gong, Lingqiao Liu, Vuong Le, Budhaditya Saha, Moussa Reda Mansour, Svetha Venkatesh, and Anton van den Hengel. Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In *Proceedings of the the IEEE/CVF International Conference on Computer Vision*, 2019.

[Goodfellow *et al.*, 2014] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Proceedings of the Advances in Neural Information Processing Systems*, 2014.

[Guan Gui, 2024] Jun Liu Chengjie Wang Yunsheng Wu Guan Gui, Bin-Bin Gao. Few-shot anomaly-driven generation for anomaly classification and segmentation. In *Proceedings of the European Conference on Computer Vision*, 2024.

[Hertz *et al.*, 2022] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022.

[Ho *et al.*, 2020] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Proceedings of the Advances in Neural Information Processing Systems*, 2020.

[Hu *et al.*, 2023] Teng Hu, Jiangning Zhang, Ran Yi, Yuzhen Du, Xu Chen, Liang Liu, Yabiao Wang, and Chengjie Wang. Anomalydiffusion: Few-shot anomaly image generation with diffusion model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023.

[Kim *et al.*, 2024] Jae Myung Kim, Jessica Bader, Stephan Alaniz, Cordelia Schmid, and Zeynep Akata. Datadream: Few-shot guided dataset generation. In *Proceedings of the European Conference on Computer Vision*, 2024.

[Li *et al.*, 2021] 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*, 2021.

[Lin *et al.*, 2021] Dongyun Lin, Yanpeng Cao, Wenbin Zhu, and Yiqun Li. Few-shot defect segmentation leveraging abundant defect-free training samples through normal background regularization and crop-and-paste operation. In *2021 IEEE International Conference on Multimedia and Expo*, 2021.

[Liu *et al.*, 2023] 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*, 2023.

[Marcos-Manchón *et al.*, 2024] Pablo Marcos-Manchón, Roberto Alcover-Couso, Juan C SanMiguel, and Jose M Martínez. Open-vocabulary attention maps with token optimization for semantic segmentation in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[Mokady *et al.*, 2023] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[Nguyen *et al.*, 2024] Quang Nguyen, Truong Vu, Anh Tran, and Khoi Nguyen. Dataset diffusion: Diffusion-based synthetic data generation for pixel-level semantic segmentation. In *Proceedings of the Advances in Neural Information Processing Systems*, 2024.

[Niu *et al.*, 2020] Shuanlong Niu, Bin Li, Xinggang Wang, and Hui Lin. Defect image sample generation with gan for improving defect recognition. *IEEE Transactions on Automation Science and Engineering*, 17(3):1611–1622, 2020.

[Otsu, 1979] Nobuyuki Otsu. A threshold selection method from gray-level histograms. *IEEE transactions on systems, man, and cybernetics*, 9(1):62–66, 1979.

[Park *et al.*, 2020] Hyunjong Park, Jongyoun Noh, and Bumsub Ham. Learning memory-guided normality for anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[Qiu *et al.*, 2025] Kunpeng Qiu, Zhiqiang Gao, Zhiying Zhou, Mingjie Sun, and Yongxin Guo. Noise-consistent siamese-diffusion for medical image synthesis and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

[Ristea *et al.*, 2022] Nicolae-Cătălin Ristea, Neelu Madan, Radu Tudor Ionescu, Kamal Nasrollahi, Fahad Shahbaz Khan, Thomas B Moeslund, and Mubarak Shah. Self-supervised predictive convolutional attentive block for anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[Rombach *et al.*, 2022] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[Roth *et al.*, 2022] 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*, 2022.

[Schlüter *et al.*, 2022] Hannah M Schlüter, Jeremy Tan, Benjamin Hou, and Bernhard Kainz. Natural synthetic anomalies for self-supervised anomaly detection and localization. In *Proceedings of the European Conference on Computer Vision*. Springer, 2022.

[Song and Ermon, 2019] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *Proceedings of the Advances in Neural Information Processing Systems*, 2019.

[Xie *et al.*, 2021] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. In *Proceedings of the Advances in Neural Information Processing Systems*, 2021.

[Yan *et al.*, 2021] 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*, 2021.

[Yao *et al.*, 2023] Xincheng Yao, Ruoqi Li, Jing Zhang, Jun Sun, and Chongyang Zhang. Explicit boundary guided semi-push-pull contrastive learning for supervised anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[You *et al.*, 2022] Zhiyuan You, Lei Cui, Yujun Shen, Kai Yang, Xin Lu, Yu Zheng, and Xinyi Le. A unified model for multi-class anomaly detection. In *Proceedings of the Advances in Neural Information Processing Systems*, 2022.

[Zavrtanik *et al.*, 2021] 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*, 2021.

[Zhang *et al.*, 2021] Gongjie Zhang, Kaiwen Cui, Tzu-Yi Hung, and Shijian Lu. Defect-gan: High-fidelity defect synthesis for automated defect inspection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021.

[Zhang *et al.*, 2023a] Hui Zhang, Zuxuan Wu, Zheng Wang, Zhineng Chen, and Yu-Gang Jiang. Prototypical residual networks for anomaly detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[Zhang *et al.*, 2023b] Manlin Zhang, Jie Wu, Yuxi Ren, Ming Li, Jie Qin, Xuefeng Xiao, Wei Liu, Rui Wang, Min Zheng, and Andy J Ma. Diffusionengine: Diffusion model is scalable data engine for object detection. *arXiv preprint arXiv:2309.03893*, 2023.

[Zhang *et al.*, 2023c] 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*, 2023.

[Zhao *et al.*, 2023] Yuzhong Zhao, Qixiang Ye, Weijia Wu, Chunhua Shen, and Fang Wan. Generative prompt model for weakly supervised object localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.