

# Dual-Interrelated Diffusion Model for Few-Shot Anomaly Image Generation

Ying Jin<sup>1\*</sup>, Jinlong Peng<sup>2\*</sup>, Qingdong He<sup>2\*</sup>, Teng Hu<sup>3</sup>, Jiafu Wu<sup>2</sup>, Hao Chen<sup>1</sup>

Haoxuan Wang<sup>1</sup>, Wenbing Zhu<sup>1</sup>, Mingmin Chi<sup>1†</sup>, Jun Liu<sup>2</sup>, Yabiao Wang<sup>2,4†</sup>

<sup>1</sup>Fudan University, <sup>2</sup>You Lab, Tencent, <sup>3</sup>Shanghai Jiao Tong University, <sup>4</sup>Zhejiang University

{yjin22, haochen22, hxiang23, wbzhu23}@m.fudan.edu.cn, mmchi@fudan.edu.cn

{jeromepeng, yingcaihe, jiafwu, juliusliu}@tencent.com, hu-teng@sjtu.edu.cn, yabiaowang@zju.edu.cn

<https://github.com/yinyjin/DualAnoDiff>

## Abstract

The performance of anomaly inspection in industrial manufacturing is constrained by the scarcity of anomaly data. To overcome this challenge, researchers have started employing anomaly generation approaches to augment the anomaly dataset. However, existing anomaly generation methods suffer from limited diversity in the generated anomalies and struggle to achieve a seamless blending of this anomaly with the original image. Moreover, the generated mask is usually not aligned with the generated anomaly. In this paper, we overcome these challenges from a new perspective, simultaneously generating a pair of the overall image and the corresponding anomaly part. We propose DualAnoDiff, a novel diffusion-based few-shot anomaly image generation model, which can generate diverse and realistic anomaly images by using a dual-interrelated diffusion model, where one of them is employed to generate the whole image while the other one generates the anomaly part. Moreover, we extract background and shape information to mitigate the distortion and bluriness phenomenon in few-shot image generation. Extensive experiments demonstrate the superiority of our proposed model over state-of-the-art methods in terms of diversity, realism and the accuracy of mask. Overall, our approach significantly improves the performance of downstream anomaly inspection tasks, including anomaly detection, anomaly localization, and anomaly classification tasks. Code will be made available.

## 1. Introduction

Industrial anomaly inspection, i.e., anomaly detection, localization, and classification, plays an important role in in-

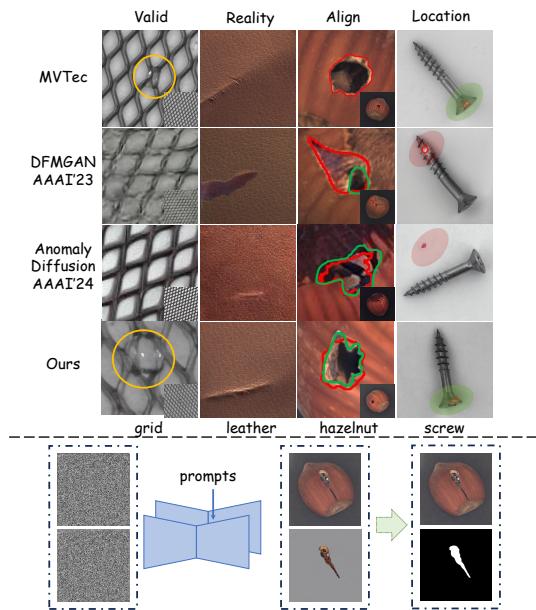


Figure 1. Top: Evaluating anomaly generation quality in four aspects: Whether generate valid anomaly, degree of realism, alignment of the mask, and whether the location of mask is reasonable, the results show that our generated results are better than the other methods. (Yellow area represents the valid generation, green indicates the correct mask or area, red indicates the generated mask or wrong area.) Bottom: Our model can simultaneously generate extensive anomaly image-mask pairs.

dustrial manufacture [6]. However, in real-world industrial production, anomaly samples are scarce. Therefore, the current mainstream anomaly inspection methods are either unsupervised methods which use only normal samples [22, 36] or semi-supervised method [48] which employ both the normal samples and a few anomaly data. Although these methods perform well in anomaly detection, they have limited performance in anomaly localization and can not deal with the task of anomaly classification [14]. Therefore, researchers proposed anomaly generation methods to gener-

\*Equal contribution.

†Corresponding author (This work was supported by Natural Science Foundation of China under contract 62171139).

ate more anomaly data, to help achieve better performance by using supervised anomaly inspection.

Existing anomaly generation methods can be categorized into two groups, 1) *model-free methods* randomly cut and paste patches from existing anomalies or anomalous texture datasets onto normal samples [20, 23, 46]. But the anomaly data synthesized by them are unrealistic. 2) *Generative methods* employ generative models like GANs and diffusion models to generate anomaly data. Generally, GAN-based model [26, 47] requires a large amount of training data to achieve better generation performance and they can not generate masks. DFMGAN [6] is firstly trained on normal data and then migrated to anomaly data to achieve a few-shot generation. This method also encounters the problem that the generated anomalies are not realistic enough [29] and the masks are not sufficiently aligned [21] because there is no explicit alignment constraint design. Anomaly-Diffusion [14] based on the texture-inversion [7] technique of Diffusion [33], separately learns the anomaly appearance and location information, then generates the anomaly on the masked normal samples. As AnomalyDiffusion only focuses on the part of anomaly, the generated anomalies do not blend realistically with the original image, and the masks generated individually may appear in the background of image. To address these limitations, we propose DualAnoDiff, a novel few-shot anomaly image generation model that utilizes a dual-interrelated diffusion to simultaneously generate the overall image and the corresponding anomaly part. This approach can realize the effective integration of anomaly image and anomaly part, resulting in the generation of realistic and highly aligned anomaly image-mask data pairs with good diversity. Specifically, our model is built upon a pre-trained diffusion model and introduces two LoRA [12] to expand a single diffusion model into **dual-interrelated diffusion model**. One branch, referred to as the global branch, is responsible for generating the overall anomaly image, while the anomalous branch generates the localized anomaly image. They exchange information through the **self-attention interaction module**. This module combines the attention layers of global and anomalous branches and performs shared attention calculations, enabling the interaction and fusion of information in the dual-denoising process. This ensures the consistency between the generated overall anomaly image and the localized anomaly image. Furthermore, to further preserve the invariance of the background, we introduce a **background compensation module** based on self-attention adaptive injection. This module involves adding noise to the background image, extracting the key and value from the intermediate feature layer, and applying adaptive fusion MLP to incorporate the background information into the global branch. It contributes to maintaining the accuracy of the background and the shape of the object in the generated im-

ages, while avoiding the coupling between the object and the background in the images.

Fig.7 shows our generated results outperform those of other methods in four key aspects. Moreover, extensive experiments have been conducted on MVTec AD [1] to quantitatively validate the superiority of the anomaly data generated by DualAnoDiff in downstream anomaly inspection tasks, and achieving a state-of-the-art level of performance in pixel-level anomaly detection with **99.1% AUROC** and **84.5% AP** score.

Our contributions can be summarized as follows:

- We propose DualAnoDiff, a novel few-shot diffusion-based anomaly generation method, which simultaneously generates both the overall image and the corresponding anomaly part with a highly aligned mask by a dual-interrelated diffusion model.
- We design a background compensation approach, which involves image backgrounds as control information and injects the intermediate feature into the denoising process of anomaly image, to enhance the stability and realism of the generated data.
- Extensive experiments demonstrate the superiority of our method over existing anomaly generation models in terms of both generation quality and performance of downstream anomaly inspection tasks.

## 2. Related Work

### 2.1. Few-shot Image Generation

Few-shot image generation aims to generate new and diverse examples while preventing overfitting to the few training images [45, 51]. It is highly susceptible to overfit with extremely limited training data (less than 10) and then generate highly similar images. FreezeD [25] proposes modifying network weights, using various regularization techniques and data augmentation to prevent overfitting [14]. There are also methods [13] to mitigate overfitting by pre-training on the source domain and subsequently migrating to fewer samples through cross-domain consistency losses to keep the generated distribution. Textual Inversion [7] and Dreambooth [37] encode a few images into the textual space of a pre-trained diffusion model to achieve diverse target customization generation which preserves its key visual features. Although these methods can generate realistic images, they are incapable of generating pixel-level annotations which are essential for anomaly image generation tasks. In contrast, our method enables high-quality annotations to be readily obtained by simultaneously generating local anomaly images.

### 2.2. Anomaly Inspection

The anomaly inspection task consists of anomaly detection, localization, and classification [6]. Due to the scarcity

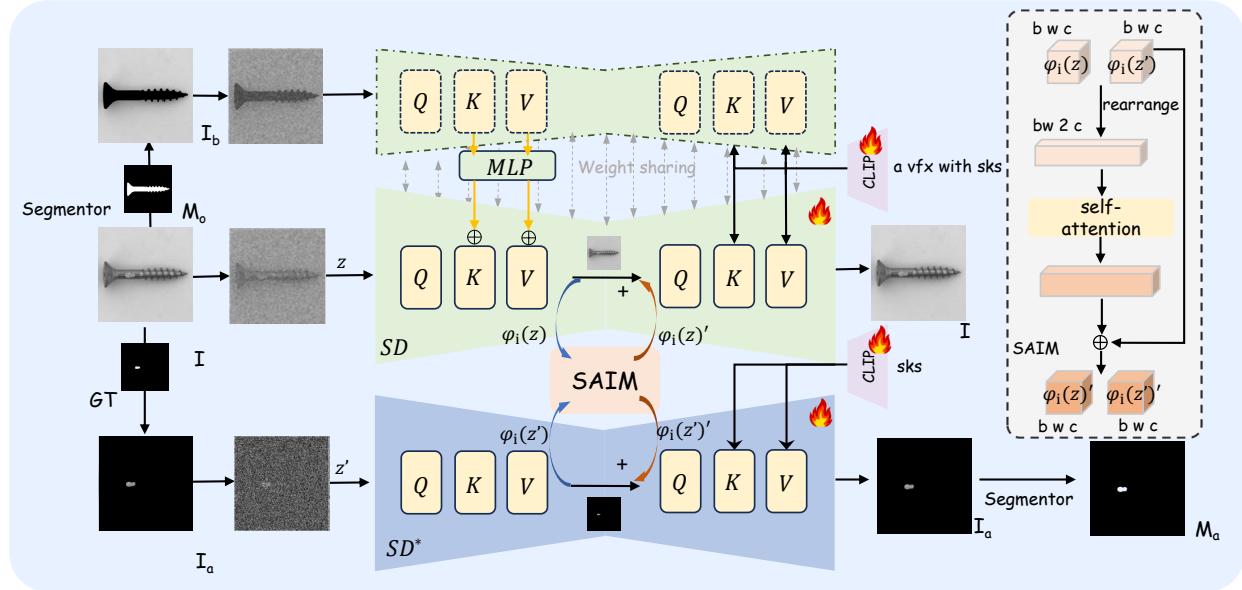


Figure 2. The architecture of DualAnoDiff. 1) Two branches of DualAnoDiff generate the anomaly image and corresponding anomaly part simultaneously with different but nested prompts. 2) Two branches share the attention information after every attention block by Self-Attention Interaction Module (SAIM) during the denoising process to keep the consistency of generated images. 3) Background Compensation Module (BCM) extracts the Key, and Value of the background image and applies an adaptive fusion to SD, to help the model more focus on the object of the image.

of abnormal data in industrial scenarios, most methods [9, 18, 22, 24, 36, 40–42] use unsupervised methods and semi-supervised methods. Reconstruction-based methods [3, 10, 39] detect anomalies by analyzing the residual image before and after reconstruction. Embedding-based methods [2, 19, 44] utilize pre-trained networks to extract the image-level features and patch-level features, and then perform clustering according to the similarity between the features to detect the anomalies. All of those methods can only address the task of anomaly detection, while having limited performance in anomaly localization and being incapable of anomaly classification. Through the generation of abnormal images, these three tasks can be successfully accomplished, and our method achieves the state-of-the-art performance.

### 2.3. Anomaly Generation

Due to the scarcity of anomaly data, anomaly generation has emerged as a field of crucial significance. DRAEM [46], Cut-Paste [20], Crop-Paste [23] and PRN [48] crop and paste unrelated textures or existing anomalies into normal sample. These approaches can be somewhat effective, but the generated anomalies are completely unrealistic and there is no way to use them as anomaly classification tasks. Subsequently, GANs [8] have been applied for anomaly generation due to their ability to generate high-fidelity images. SDGAN [26] and DefectGAN [47] generate anomalies on normal samples by learning from anomaly

data. However, they require a large amount of anomaly data and cannot generate anomaly masks. DFMGAN [6] transfers a StyleGAN2 [15] pretrained on normal samples to the anomaly domain, but lacks generation realism and accurate alignment between generated anomalies and masks. Subsequently, Diffusion models have been more widely used due to their extreme generalizability. AnomalyDiffusion [14] learns the features of anomaly and the distribution of masks through text inversion [7] technique of diffusion, to generate the specified anomaly at the position of the corresponding mask of the normal image. However, Since the method learns the anomaly part and the mask separately, it makes the generated mask not necessarily located on the object, and the anomaly-object transition is not natural enough. Nevertheless, through the utilization of a two-branch diffusion model along with a background supplement module, our methodology successfully accomplishes the full decoupling of diverse attributes within the anomaly image. As a result, the generated data are more realistic and diversified, and the performance of various downstream tasks is significantly enhanced.

## 3. Method

Given a set of a limited number of anomaly image-mask pairs, our goal is to learn the features of the anomaly images, and then generate more anomaly image-mask pairs

that belong to the same item and anomaly type, while ensuring greater diversity in the distribution and appearance of the anomaly. In our method, we generate the overall image and the anomaly part, which is then segmented to get the corresponding mask.

### 3.1. Preliminaries

**Latent Diffusion Models.** Stable Diffusion (SD), a variant of the latent diffusion model (LDM) [33], serves as a text-guided diffusion model. To generate high-resolution images while enhancing computational efficiency in the training process, it employs a pre-trained variational autoencoder (VAE) [16] encoder  $\mathcal{E}(\cdot)$  to map images into latent space and perform an iterative denoising process. Subsequently, the predicted images are mapped back into pixel space through the pre-trained VAE decoder  $D(\cdot)$ .  $\epsilon_\theta$  is the denoising network, for each denoising step, the simplified optimization objective is defined as follows:

$$L_{LDM}(\theta) = \mathbb{E}_{\mathcal{E}(x), \epsilon, t} [\|\epsilon - \epsilon_\theta(z_t, t, \tau_\theta(c))\|_2^2] \quad (1)$$

where  $\epsilon$  are latent noise, the text description  $c$  is encoded by the CLIP [31] text encoder  $\tau_\theta(\cdot)$  and then used to guide the diffusion denoising process.

### 3.2. DualAnoDiff Framework

We have been searching for methods that can generate both anomaly image and mask. Inspired by Layerdiffusion [49], we decompose an anomaly image into two parts, the overall anomaly image and the corresponding anomaly part, where the overall image refers to the whole anomaly image  $I$ , the anomaly part refers to the part that contains only the anomaly region  $I_a$ .  $I_a = I \times M_a$  ( $M_a$  is the mask of anomaly part).

As shown in Fig. 2, the proposed *DualAnoDiff* involves two interrelated diffusion models, and they share part information by the Self-attention Interaction Module. We freeze the weights of diffusion models and use two LORA [12] to fine-tune them. For ease of description, we denote the two diffusion models as *SD* and *SD\**, *SD* denote the diffusion model to generate global image  $I$ , *SD\** denote the diffusion model to generate anomaly part  $I_a$ .

**Dual-Interrelated Diffusion.** AnomalyDiffusion [14] primarily focuses on the anomaly part, which may result in generated anomaly images lacking a convincingly realistic appearance. However, generating the complete anomaly image poses challenges in obtaining the corresponding mask. To address those limitations, our proposed model simultaneously generates both the overall image and the anomaly part. This novel approach overcomes challenges in generating realistic anomaly images while ensuring the availability of accurate masks.

First, we encode  $I$ ,  $I_a$  into latent space  $z$  and  $z'$ , with  $z = \varepsilon(I)$  and  $z' = \varepsilon(I_a)$  by using VAE encoder  $\varepsilon(\cdot)$ .

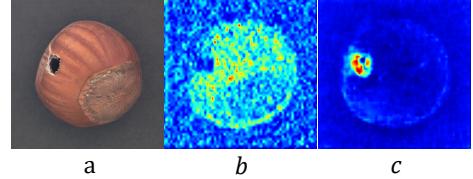


Figure 3. a is the image generated by SD, b and c are the cross attention maps of different text tokens in SD corresponding to the text of “a vfx with” and “sks”.

Next we employ a forward process to add noise into the latents with the same timestep  $t$ , and then learn to denoising during the backward process guided by different prompts. Throughout these processes, information is shared and synchronized between the two diffusion models through the SAIM, enabling the model to effectively fit the training data pairs. By generating the anomalies separately, this approach achieves two important objectives. Through a simple yet effective operation of adding two LoRA, it enhances the diversity and reality of the generated anomaly. Additionally, it ensures a highly aligned mask that accurately corresponds to the anomaly image.

**Nested Prompts.** The goal of the dual diffusion is to generate an anomaly image and anomaly part pair ( $I$  and  $I_a$ ), which exhibits an inclusion relationship. To facilitate the model’s understanding of the distinct entities within the image (the primary subject and the anomaly), we employ a pair of prompts designed to reflect this inclusion relationship:

$$\begin{aligned} p &: a \text{ } x \text{ with } y \\ p' &: y \end{aligned} \quad (2)$$

where the prompts  $p$  and  $p'$  correspond to the anomaly image  $I$  and the corresponding anomaly part  $I_a$  respectively. Both prompts are encoded by the trainable text encoder  $\tau_\theta(\cdot)$  and then injected into Unet [34] of Diffusion.

The variables  $x$  and  $y$  can be the class name and anomaly name provided by the dataset. In our model, we use the *vfx* and *sks* which were suggested by DreamBooth [37]. Those words have weak prior in both the language model and the diffusion model, making them easier to fit than other words, and can achieve better generation results, specially for high-prior words. Fig.3 presents the generated result and visualizes the cross-attention maps between text token and vision. where  $a$  is the anomaly image generated by *SD*.  $b$  and  $c$  are the  $64 \times 64$  resolutions feature maps randomly extracted from the second half of the generation process in the Unet of *SD*. Where  $b$  corresponds to the text “*a vfx with*”,  $c$  corresponds to “*sks*”. It is evident that the model correctly separates the attributes of anomaly and object, and accurately associating them to the specified text as we want.

**Self-attention Interaction Module (SAIM).** During training, *SD* and *SD\** share the same timestep  $t$  and denois-



Figure 4. Comparison between the models without (a) and with (b) the Background Compensation Module.

ing simultaneously, they share information by SAIM after every attention blocks in the Unet of diffusion. For example, after the self-attention blocks is more likely to share the positional information and detailed information, the cross-attention blocks shares the semantic information.

In SAIM, we use attention to fuse the information from two branch, The shared step is formulated as:

$$\begin{aligned}\varphi_i(\tilde{z}) &= \text{Rearrange}(\text{Concat}(\varphi_i(z), \varphi_i(z'))) \\ \varphi_i(\tilde{z})_{\text{new}} &= \text{SelfAtt}(\varphi_i(\tilde{z})) \\ \varphi_i(z'), \varphi_i(z')' &= \text{Split}(\text{Rearrange}(\varphi_i(\tilde{z})_{\text{new}} + \varphi_i(\tilde{z})))\end{aligned}\quad (3)$$

where  $\varphi_i$  is the intermediate representation of the Unet. The original shape of  $\varphi_i(z)$  and  $\varphi_i(z')$  are “b w c”, where “b” denotes the batch size, “w” represents the spatial dimension, and “c” represents the channel dimension. We rearrange the  $\text{Concat}(\varphi_i(z), \varphi_i(z'))$  to shape of “bw 2 c” to avoid displacement in space.  $\text{Concat}$  and  $\text{Split}$  are a set of corresponding operations that are used to aggregate and separate feature maps.

**Loss Function.** With the two-stream structure of diffusion, our final training objective is expressed as follows:

$$\begin{aligned}\mathcal{L} = &\mathbb{E}_{\mathcal{E}(I), \epsilon, t} [\|\epsilon - \epsilon_\theta(z_t, t, \tau_\theta(p))\|_2^2] \\ &+ \mathbb{E}_{\mathcal{E}(I_a), \epsilon^*, t} [\|\epsilon^* - \epsilon_\theta^*(z'_t, t, \tau_\theta(p'))\|_2^2]\end{aligned}\quad (4)$$

where,  $\epsilon$  and  $\epsilon^*$  are latent noise for the anomaly image and anomaly part,  $t$  is the time step,  $p, p'$  are the corresponding prompts, they are encoded by text encoder  $\tau_\theta(\cdot)$  which is trainable.

**Mask Generation.** Thanks to our two-stream parallel structure, which generates the anomaly part as a single entity, obtaining the precise mask becomes straightforward. There are two ways to obtain a high-quality mask: 1). We utilize existing segmentation algorithms such as SAM [17], U<sup>2</sup>-Net [30] to obtain high-quality masks after generating the anomaly part image. 2) During the generation process, we can extract the average attention maps in  $SD^*$  to compute the mask such as [43] which is widely used in semantic segmentation. In this paper, we use the first method.

### 3.3. Background Compensation Module

Although the model has achieved good results so far, there are still some problems in some cases, which is caused by the limited training data. Several bad cases of corresponding generated results are shown in Fig.4 for categories bottle, pill, and toothbrush in the MVTec (the average training data for each category is 8). The bottle was only partially generated. The edges of pill is blurred and the internal properties leakage into the background. Additionally, the toothbrush has two brush handles, which is an abnormal occurrence. Furthermore, all of those generated images lack sharpness and the background color does not maintain sufficient purity. In general, since the model is incapable of fully grasping the characteristics of those cases, there are problems such as objects being mixed with the background and objects being deformed.

To enhance the model’s ability to learn the shape of object and focus more on object generation, we design a background compensation module. First, we employ U<sup>2</sup>-Net to segment the image  $I$  and obtain the object mask  $M_f$ , then get the background image  $I_b = (1 - M_f) \times I$ ,  $I_b$  will be processed by the  $SD$  as same as  $I$  (except for the SAIM). Then, we utilize  $I_b$  as a condition, which contains both background and mask information. This allows us to control the shape of the object while providing background context. We collect its features in self-attentions and inject them into the  $K, V$  of the  $I$  in every corresponding self-attention steps. The injection process can be formulated as:

$$\begin{aligned}\varphi_i(z^b) &= \text{SelfAtt}(\varphi_i(z^b)) \\ Q &= W_Q^{(i)} \cdot \varphi_i(z) \\ \varphi_i(z) &= \varphi_i(z) + \gamma \text{MLP}(\varphi_i(z^b)) \\ K &= W_K^{(i)} \cdot \varphi_i(z) \quad V = W_V^{(i)} \cdot \varphi_i(z)\end{aligned}\quad (5)$$

where  $\gamma$  is a learnable scale factor initialized to be 0.1. This design can help us to maximally preserve the generative effect of the mainstream SD, meanwhile use background information and the shape of the mask. With the condition of  $I_b$ , the loss function becomes:

$$\begin{aligned}\mathcal{L} = &\mathbb{E}_{\mathcal{E}(I), \epsilon, t} [\|\epsilon - \epsilon_\theta(z_t, t, \tau_\theta(p), I_b)\|_2^2] \\ &+ \mathbb{E}_{\mathcal{E}(I_a), \epsilon^*, t} [\|\epsilon^* - \epsilon_\theta^*(z'_t, t, \tau_\theta(p'), I_b)\|_2^2]\end{aligned}\quad (6)$$

From another perspective, the dual-branch structure is equivalent to explicitly separating the two concepts of the object body and the anomaly in image. However, in the case of some images with backgrounds, two attributes are not enough. model may mix the object body with the solid-color background, thus resulting in the phenomenon shown in Fig. 4(a). The BCM module further decouples the object

Category	CDC		Crop-Paste		SDGAN		Defect-GAN		DFMGAN		AnomalyDiffusion		Ours	
	IS ↑	IC-L ↑	IS ↑	IC-L ↑	IS ↑	IC-L ↑	IS ↑	IC-L	IS ↑	IC-L ↑	IS ↑	IC-L ↑	IS ↑	IC-L ↑
bottle	1.52	0.04	1.43	0.04	1.57	0.06	1.39	0.07	<u>1.62</u>	0.12	1.58	0.19	<b>2.17</b>	<b>0.43</b>
cable	1.97	0.19	1.74	0.25	1.89	0.19	1.70	0.22	1.96	0.25	<u>2.13</u>	0.41	<b>2.15</b>	<b>0.43</b>
capsule	1.37	0.06	1.23	0.05	1.49	0.03	<u>1.59</u>	0.04	<u>1.59</u>	0.11	<u>1.59</u>	0.21	<b>1.62</b>	<b>0.32</b>
carpet	<u>1.25</u>	0.03	1.17	0.11	1.18	0.11	1.24	0.12	1.23	0.13	1.16	0.24	<b>1.36</b>	<b>0.29</b>
grid	1.97	0.07	2.00	0.12	1.95	0.10	2.01	0.12	1.97	0.13	<u>2.04</u>	<b>0.44</b>	<b>2.13</b>	0.42
hazel_nut	<u>1.97</u>	0.05	1.74	0.21	1.85	0.16	1.87	0.19	1.93	0.24	<b>2.13</b>	0.31	1.94	<b>0.35</b>
leather	1.80	0.07	1.47	0.14	2.04	0.12	<b>2.12</b>	0.14	<u>2.06</u>	0.17	1.94	<b>0.41</b>	1.91	0.35
metal_nut	1.55	0.04	<u>1.56</u>	0.15	1.45	0.28	1.47	<u>0.30</u>	1.49	<b>0.32</b>	<b>1.96</b>	0.30	<u>1.57</u>	<b>0.32</b>
pill	1.56	0.06	1.49	0.11	1.61	0.07	1.61	0.10	<u>1.63</u>	0.16	1.61	<u>0.26</u>	<b>1.82</b>	<b>0.38</b>
screw	1.13	0.11	1.12	0.16	1.17	0.10	1.19	0.12	1.12	0.14	<u>1.28</u>	0.30	<b>1.43</b>	<b>0.36</b>
tile	2.10	0.12	1.83	0.20	<u>2.53</u>	0.21	2.35	0.22	2.39	0.22	<b>2.54</b>	<b>0.55</b>	2.40	0.50
toothbrush	1.63	0.06	1.30	0.08	1.78	0.03	<u>1.85</u>	0.03	1.82	0.18	1.68	<u>0.21</u>	<b>2.40</b>	<b>0.48</b>
transistor	1.61	0.13	1.39	0.15	<b>1.76</b>	0.13	1.47	0.13	1.64	0.25	1.57	<b>0.34</b>	1.71	0.33
wood	2.05	0.03	1.95	0.23	2.12	0.25	2.19	0.29	2.12	0.35	<b>2.33</b>	0.37	2.24	<b>0.40</b>
zipper	1.30	0.05	1.23	0.11	1.25	0.10	1.25	0.10	1.29	<u>0.27</u>	<u>1.39</u>	0.25	<b>2.14</b>	<b>0.37</b>
Average	1.65	0.07	1.51	0.14	1.71	0.13	1.69	0.15	1.72	0.20	1.80	0.32	<b>1.93</b>	<b>0.38</b>

Table 1. Comparison on IS and IC-LPIPS on MVTec dataset. Our model generates the most high-quality and diverse anomaly data, achieving the best IS and IC-LPIPS. Bold and underline represent optimal and sub-optimal results, respectively.

body from the background by incorporating background information during training, thereby achieving a more stable generation effect in Fig. 4(b).

## 4. Experiments

### 4.1. Experiment Settings

**Dataset.** We conduct experiments on MVTec [1] and follow the experimental setup of AnomalyDiffusion [14] using 1/3 of the data for training, reserving the remaining 2/3 for testing.

**Implementation Details.** We train a model for each anomaly type separately and generate 1000 anomaly image-mask pairs for the downstream anomaly inspection tasks. More details of the experiment can be found in the supplementary materials.

**Metric.** 1) **For generation**, we use Inception Score (IS) and Intra-cluster pairwise LPIPS distance (IC-LPIPS) [27] as AnomalyDiffusion [14] to measure the generation quality and generation diversity. 2) **For anomaly inspection**. We use AUROC, Average Precision (AP), and the F1-max score to evaluate the accuracy of anomaly detection and localization, use Accuracy to evaluate the anomaly classification.

### 4.2. Comparison in Anomaly Generation

**Baseline.** 1) We choose CDC [27], Crop-Paste[23], SDGAN [26], DefectGAN [47], DFMGAN [6] and AnomalyDiffusion [14] that can generate specific anomaly types to compare anomaly generation quality and classification. 2) We select DRAEM[46], PRN [48], DFMGAN [6] and AnomalyDiffusion [14] which can generate anomaly image-mask pairs as comparative benchmarks to compare anomaly detection and localization.

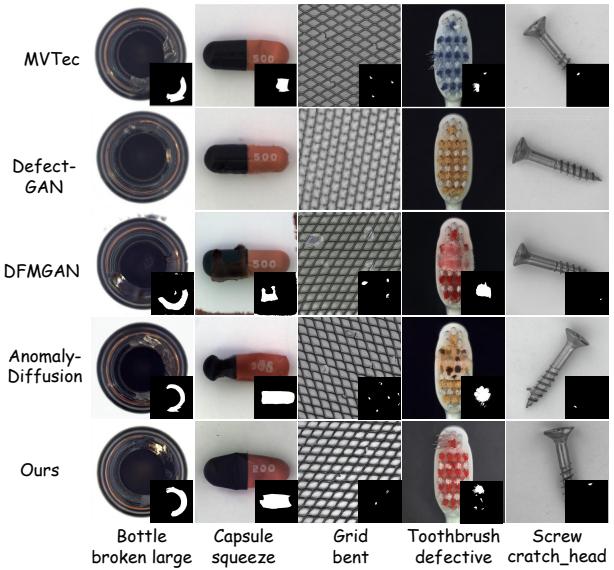


Figure 5. Comparison of the generation results on MVTec. Our model excels in generating high-quality anomaly images that are accurately aligned with the anomaly masks.

**Anomaly Generation Quality.** Tab. 1 shows the results of the quantification metrics for image quality. For each anomaly category, we allocate 1/3 of the anomaly data for training and generate 1000 anomaly images to compute IS and IC-LPIPS. Through this process, it is clearly demonstrated that our method achieves the best results in both the IS and IC-LPIPS metrics. Moreover, it further shows that our model generates anomaly data with both the highest quality and diversity.

Category	DRAEM				PRN				DFMGAN				AnomalyDiffusion				Ours			
	AUC-P	AP-P	$F_1$ -P	AP-I	AUC-P	AP-P	$F_1$ -P	AP-I	AUC-P	AP-P	$F_1$ -P	AP-I	AUC-P	AP-P	$F_1$ -P	AP-I	AUC-P	AP-P	$F_1$ -P	AP-I
bottle	96.7	80.2	74.0	99.8	97.5	76.4	71.3	98.4	98.9	90.2	83.9	99.8	99.4	<b>94.1</b>	<b>87.3</b>	<b>99.9</b>	<b>99.5</b>	93.4	85.7	<b>100</b>
cable	80.3	21.8	28.3	83.2	94.5	64.4	61.0	92.0	97.2	81.0	75.4	97.8	<b>99.2</b>	<b>90.8</b>	<b>83.5</b>	<b>100</b>	97.5	82.6	76.9	98.3
capsule	76.2	25.5	32.1	98.7	95.6	45.7	47.9	95.8	79.2	26.0	35.0	98.5	<b>98.8</b>	<b>57.2</b>	<b>59.8</b>	<b>99.9</b>	<b>99.5</b>	<b>73.2</b>	<b>67.0</b>	<b>99.2</b>
carpet	92.6	43.0	41.9	98.7	96.4	69.6	65.6	97.8	90.6	33.4	38.1	98.5	<b>98.6</b>	<b>81.2</b>	<b>74.6</b>	<b>98.8</b>	<b>99.4</b>	<b>89.1</b>	<b>80.2</b>	<b>99.9</b>
grid	<b>99.1</b>	<b>59.3</b>	<b>58.7</b>	<b>99.9</b>	<b>98.9</b>	<b>58.6</b>	<b>58.9</b>	<b>98.9</b>	75.2	14.3	20.5	90.4	98.3	52.9	54.6	99.5	98.5	57.2	54.9	<b>99.7</b>
hazelnut	98.8	73.6	68.5	<b>100</b>	98.0	73.9	68.2	96.0	<b>99.7</b>	95.2	89.5	<b>100</b>	<b>99.8</b>	<b>96.5</b>	<b>90.6</b>	<b>99.9</b>	<b>99.8</b>	<b>97.7</b>	<b>92.8</b>	<b>100</b>
leather	98.5	67.6	65.0	<b>100</b>	99.4	58.1	54.0	<b>99.7</b>	98.5	68.7	66.7	<b>100</b>	<b>99.8</b>	<b>79.6</b>	<b>71.0</b>	<b>100</b>	<b>99.9</b>	<b>88.8</b>	<b>78.8</b>	<b>100</b>
metal_nut	96.9	84.2	74.5	99.6	97.9	93.0	87.1	99.5	99.3	<b>98.1</b>	<b>94.5</b>	99.8	<b>99.8</b>	<b>98.7</b>	<b>94.0</b>	<b>100</b>	<b>99.6</b>	98.0	93.0	<b>99.9</b>
pill	95.8	45.3	53.0	98.9	98.3	55.5	72.6	97.8	81.2	67.8	72.6	91.7	<b>99.8</b>	<b>97.0</b>	<b>90.8</b>	<b>99.6</b>	<b>99.6</b>	<b>95.8</b>	89.2	99.0
screw	91.0	30.1	35.7	<b>96.3</b>	94.0	47.7	49.8	94.7	58.8	2.2	5.3	64.7	<b>97.0</b>	<b>51.8</b>	<b>50.9</b>	<b>97.9</b>	<b>98.1</b>	<b>57.1</b>	<b>56.1</b>	95.0
tile	98.5	93.2	87.8	<b>100</b>	98.5	91.8	84.4	<b>96.9</b>	<b>99.5</b>	<b>97.1</b>	<b>91.6</b>	<b>100</b>	99.2	<b>93.9</b>	86.2	<b>100</b>	<b>99.7</b>	<b>97.1</b>	<b>91.0</b>	<b>100</b>
toothbrush	93.8	29.5	28.4	99.8	96.1	46.4	46.2	<b>100</b>	96.4	<b>75.9</b>	<b>72.6</b>	<b>100</b>	<b>99.2</b>	<b>76.5</b>	<b>73.4</b>	<b>100</b>	<b>98.2</b>	68.3	68.6	99.7
transistor	76.5	31.7	24.2	80.5	94.9	68.6	68.4	88.9	96.2	81.2	77.0	92.5	<b>99.3</b>	<b>92.6</b>	<b>85.7</b>	<b>100</b>	<b>98.0</b>	86.7	79.6	<b>93.7</b>
wood	98.8	<b>87.8</b>	<b>80.9</b>	<b>100</b>	96.2	74.2	67.4	92.7	95.3	70.7	65.8	99.4	98.9	84.6	74.5	99.4	<b>99.4</b>	<b>91.6</b>	<b>83.8</b>	<b>99.9</b>
zipper	93.4	65.4	64.7	<b>100</b>	98.4	79.0	73.7	99.7	92.9	65.6	64.9	<b>99.9</b>	99.4	86.0	79.2	<b>100</b>	<b>99.6</b>	<b>90.7</b>	<b>82.7</b>	<b>100</b>
Average	92.2	54.1	53.1	97.0	96.9	66.2	64.7	96.6	90.0	62.7	62.1	94.8	<b>99.1</b>	<b>81.4</b>	<b>76.3</b>	<b>99.7</b>	<b>99.1</b>	<b>84.5</b>	<b>78.8</b>	98.9

Table 2. **Quantitative result for anomaly detection.** Comparison on pixel-level anomaly localization and image-level anomaly detection on MVTec dataset by training an U-Net on the generated data from DRAEM, PRN, DFMGAN, AnomalyDiffusion and our model.

Category	DiffAug	Crop-Paste	DFMGAN	AnoDiff	Ours
bottle	48.84	52.71	56.59	<b>90.70</b>	79.07
cable	21.36	32.81	45.31	<b>67.19</b>	<b>78.12</b>
capsule	34.67	32.8	37.23	<b>66.67</b>	<b>70.67</b>
carpet	35.48	27.96	47.31	<b>58.06</b>	<b>79.03</b>
grid	28.33	28.33	40.83	42.50	<b>80.0</b>
hazelnut	65.28	59.03	81.94	<b>85.42</b>	<b>89.58</b>
leather	40.74	34.39	49.73	<b>61.90</b>	<b>90.48</b>
metalnut	58.85	59.89	64.58	59.38	<b>89.06</b>
pill	29.86	26.74	29.52	<b>59.38</b>	56.25
screw	25.10	28.81	37.45	<b>48.15</b>	<b>70.37</b>
tile	59.65	68.42	74.85	<b>84.21</b>	<b>100</b>
transistor	38.09	41.67	52.38	<b>60.71</b>	<b>71.43</b>
wood	41.27	47.62	49.21	<b>71.43</b>	<b>85.71</b>
zipper	22.76	26.42	27.64	<b>69.51</b>	<b>75.61</b>
Average	39.31	40.55	49.61	<b>66.09</b>	<b>79.67</b>

Table 3. **Quantitative result for anomaly classification.** Comparison on the classification accuracy(%) trained on the generated data by the anomaly generation models with a ResNet-18. The higher classification accuracy indicates that the generated data is more consistent with the distribution of real data.

Moreover, we present the anomaly images generated by several prominent anomaly image generation models in Fig. 5. It is observed that the images generated by Defect-GAN often do not have anomalies, DFMGAN tends to introduce noise in the image (bottle, capsule), and the anomaly image-mask pairs lack proper alignment (grid). The state-of-the-art model AnomalyDiffusion sometimes generates unreasonable masks outside the objects (grid, screw), and the generated anomaly does not fit the original image enough to make the anomaly look unreal (toothbrush). In contrast, our proposed model demonstrates the ability to generate highly realistic and valid anomalies, even excelling in handling smaller-scale anomalies such as grid-bent. More results are

presented in the supplementary material.

**Anomaly Generation for Anomaly Detection and Localization.** We evaluate the effectiveness of our approach by comparing it with existing methods for anomaly generation in downstream anomaly detection and localization. We calculate pixel-level dimensions and image-level AUROC, AP, and F1-max for each category. The results are presented in Tab. 2, and because the image-level results of AUROC, AP, and F1-max are relatively similar, as well as the space limitation, we only show the results of the most representative AP, and the other results of AUROC and F1-max are shown in the supplementary material. It can be seen that the U-net trained in our generated data reaches the highest AP of **84.5%** and F1-max of **78.8%**, surpassing the second-ranked AnomalyDiffusion by a margin of **3.1% (AP)**.

**Anomaly Generation for Anomaly Classification.** To further assess the quality of the anomaly images generated by our model, we utilize them to train a downstream anomaly classification model. Following the experimental setup of DFMGAN [6], we employ ResNet-18 to train on the generated dataset. We then evaluate the classification accuracy on our test dataset as shown in Tab. 3. As evident from the data, our model demonstrates significantly higher accuracy across the majority of categories compared to other models, with an average accuracy improvement of **13.58%**, indicating the anomaly data we generate is more realistic.

### 4.3. Comparison with Anomaly Detection Models

To further validate the effectiveness of our model, we compare it with state-of-the-art anomaly detection methods DRAEM[46], SSPCAB [32], CFA [19], RD4AD [4], PatchCore [36], MuSc [22], DevNet [28], DRA [5] and PRN [48]. We utilize their official codes or pre-trained models

Category	Unsupervised						Supervised			
	DRAEM	SSPCAB	CFA	RD4AD	PatchCore	Musc	DevNet	DRA	PRN	Ours
bottle	99.1/88.5	98.9/88.6	98.9/50.9	98.8/51.0	97.6/75.0	98.5/82.8	96.7/67.9	91.7/41.5	99.4/92.3	<b>99.5/93.4</b>
cable	94.8/61.4	93.1/52.1	98.4/79.8	<b>98.8/77.0</b>	96.8/65.9	96.2/58.8	97.9/67.6	86.1/34.8	<b>98.8/78.9</b>	97.5/ <b>82.6</b>
capsule	97.6/47.9	90.4/48.7	98.9/71.1	99.0/60.5	98.6/46.6	98.9/52.7	91.1/46.6	88.5/11.0	98.5/62.2	<b>99.5/73.2</b>
carpet	96.3/62.5	92.3/49.1	99.1/47.7	<b>99.4/46.0</b>	98.7/65.0	<b>99.4/75.3</b>	94.6/19.6	98.2/54.0	99.0/82.0	<b>99.4/89.1</b>
grid	99.5/53.2	<b>99.6/58.2</b>	<b>98.6/82.9</b>	98.0/75.4	97.2/23.6	98.6/37.0	90.2/44.9	86.2/28.6	98.4/45.7	98.5/57.2
hazelnut	99.5/88.1	99.6/94.5	98.5/80.2	94.2/57.2	97.6/55.2	99.3/74.4	76.9/46.8	88.8/20.3	<b>99.7/93.8</b>	<b>99.8/97.7</b>
leather	98.8/68.5	97.2/60.3	96.2/60.9	96.6/53.5	98.9/43.4	99.7/62.6	94.3/66.2	97.2/5.1	99.7/69.7	<b>99.9/88.8</b>
metal_nut	98.7/91.6	99.3/95.1	98.6/74.6	97.3/53.8	97.5/86.8	87.5/49.6	93.3/57.4	80.3/30.6	<b>99.7/98.0</b>	99.6/ <b>98.0</b>
pill	97.7/44.8	96.5/48.1	98.8/67.9	98.4/58.1	97.0/75.9	97.6/65.6	98.9/79.9	79.6/22.1	99.5/91.3	<b>99.6/95.8</b>
screw	<b>99.7/72.9</b>	99.1/62.0	98.7/61.4	99.1/51.8	98.7/34.2	98.9/31.7	66.5/21.1	51.0/5.1	97.5/44.9	98.1/57.1
tile	99.4/96.4	99.2/96.3	98.6/92.6	97.4/78.2	94.9/56.0	98.3/80.6	88.7/63.9	91.0/54.4	99.6/96.5	<b>99.7/97.1</b>
toothbrush	97.3/49.2	97.5/38.9	98.4/61.7	99.0/63.1	97.6/37.1	99.4/64.2	96.3/52.4	74.5/4.8	<b>99.6/78.1</b>	98.2/68.3
transistor	92.2/56.0	85.3/36.5	98.6/82.9	<b>99.6/50.3</b>	91.8/66.7	92.5/61.2	55.2/4.4	79.3/11.2	98.4/85.6	98.0/ <b>86.7</b>
wood	97.6/81.6	97.2/77.1	97.6/25.6	99.3/39.1	95.7/54.3	98.7/77.5	93.1/47.9	82.9/21.0	97.8/82.6	<b>99.4/91.6</b>
zipper	98.6/73.6	98.1/78.2	95.9/53.9	<b>99.7/52.7</b>	98.5/63.1	98.4/64.2	92.4/53.1	96.8/42.3	98.8/77.6	99.6/ <b>90.7</b>
Average	97.7/69.0	96.2/65.5	98.3/66.3	98.3/57.8	97.1/56.6	97.5/62.6	86.4/49.3	84.8/25.7	99.0/78.6	<b>99.1/84.5</b>

Table 4. Quantitative result for pixel-level anomaly localization (AUROC/AP). We employ a simple U-Net trained on our generated dataset and the existing anomaly detection methods with their official codes or pre-trained models.

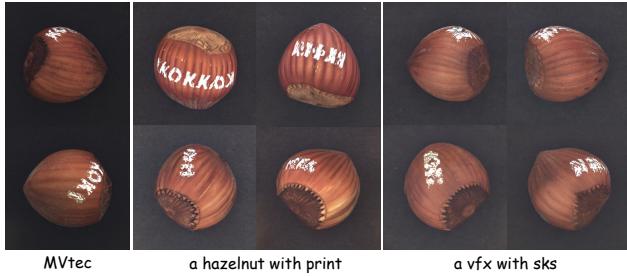


Figure 6. Comparison of generation results for different prompts on the hazelnut-print category.

and evaluate them on the same testing dataset (2/3 anomaly data in the test of MVTec). As there is no available open-source code for PRN, we rely on the results provided in its research paper. The methods are evaluated based on their pixel-level AUROC and AP scores, as demonstrated in Tab. 4. It is clearly discernible that upon using the anomaly data generated by our method, even if only the simplest U-Net model is utilized, the result shows a significant increase of **5.9%** over that of the traditional method.

#### 4.4. Ablation Study

Fig. 6 presents the results of using different prompts in our model. It is evident that using the prompt “a vfx with sks” yields superior performance in terms of object shape, color, and other aspects compared to using the category name directly. The reason is that using real object names and defect category names as prompt introduces a significant amount of prior knowledge. However, the prior knowledge from the natural world may differ greatly from the specific objects and defects in the specific dataset. More quantitative results can be found in the supplementary material.

Furthermore, we assess the effectiveness of the Back-

Category	Without BCM			With BCM		
	AUROC	AP	$F_1$ -max	AUROC	AP	$F_1$ -max
bottle	98.4	88.8	77.1	<b>99.5</b>	<b>93.4</b>	<b>85.7</b>
pill	98.4	86.9	78.2	<b>99.6</b>	<b>95.8</b>	<b>89.2</b>
toothbrush	97.2	62.7	64.0	<b>98.2</b>	<b>68.3</b>	<b>68.6</b>
Average	98.8	83.0	78.1	<b>99.1</b>	<b>84.5</b>	<b>78.8</b>

Table 5. Comparison on pixel-level anomaly localization on part of categories in MVTec between with/without the Background Compensation Module. Average is the mean of all categories.

ground Compensation Module (BCM), which exhibits more noticeable improvements in the categories of bottle, pill, and toothbrush. The visualization results are depicted in Fig. 4 and the pixel-level localization results are presented in Tab. 5. All the qualitative image generation results and quantitative metrics are improved significantly by using BCM, demonstrating BCM is critical for certain categories.

## 5. Conclusion

In this paper, we present a novel approach, DualAnoDiff, for generating anomalous image-mask pairs. Our method employs a parallel dual-diffusion to simultaneously generate the anomaly image and the corresponding anomaly part. This ensures a high level of alignment between the generated anomaly image-mask pair and the realism of the anomaly image. Additionally, to address challenging cases, we introduce a background compensation module that efficiently enhances the model’s fitting capability. Extensive experiments demonstrate its superior performance compared to existing anomaly generation methods. The anomaly data generated by our model effectively enhances the performance of downstream anomaly inspection tasks.

## References

- [1] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Mvtac ad—a comprehensive real-world dataset for unsupervised anomaly detection. In *CVPR*, pages 9592–9600, 2019. 2, 6
- [2] Yunkang Cao, Qian Wan, Weiming Shen, and Liang Gao. Informative knowledge distillation for image anomaly segmentation. *Knowledge-Based Systems*, 248:108846, 2022. 3
- [3] Yunkang Cao, Xiaohao Xu, Chen Sun, Yuqi Cheng, Zongwei Du, Liang Gao, and Weiming Shen. Segment any anomaly without training via hybrid prompt regularization. *arXiv preprint arXiv:2305.10724*, 2023. 3
- [4] Hanqiu Deng and Xingyu Li. Anomaly detection via reverse distillation from one-class embedding. In *CVPR*, pages 9737–9746, 2022. 7
- [5] Choubo Ding, Guansong Pang, and Chunhua Shen. Catching both gray and black swans: Open-set supervised anomaly detection. In *CVPR*, pages 7388–7398, 2022. 7
- [6] Yuxuan Duan, Yan Hong, Li Niu, and Liqing Zhang. Few-shot defect image generation via defect-aware feature manipulation. In *AAAI*, pages 571–578, 2023. 1, 2, 3, 6, 7
- [7] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022. 2, 3
- [8] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 3
- [9] Zhihao Gu, Jiangning Zhang, Liang Liu, Xu Chen, Jinlong Peng, Zhenye Gan, Guannan Jiang, Annan Shu, Yabiao Wang, and Lizhuang Ma. Rethinking reverse distillation for multi-modal anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8445–8453, 2024. 3
- [10] Liren He, Zhengkai Jiang, Jinlong Peng, Liang Liu, Qiang-gang Du, Xiaobin Hu, Wenbing Zhu, Mingmin Chi, Yabiao Wang, and Chengjie Wang. Learning unified reference representation for unsupervised multi-class anomaly detection. In *Proceedings of the European Conference on Computer Vision*, 2024. 3
- [11] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *NIPS*, 30, 2017. 1
- [12] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 2, 4
- [13] Teng Hu, Jiangning Zhang, Liang Liu, Ran Yi, Siqi Kou, Haokun Zhu, Xu Chen, Yabiao Wang, Chengjie Wang, and Lizhuang Ma. Phasic content fusing diffusion model with directional distribution consistency for few-shot model adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2406–2415, 2023. 2
- [14] 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*, pages 8526–8534, 2024. 1, 2, 3, 4, 6
- [15] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *CVPR*, pages 8110–8119, 2020. 3
- [16] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 4
- [17] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 5
- [18] Joo Chan Lee, Taejune Kim, Eunbyung Park, Simon S Woo, and Jong Hwan Ko. Continuous memory representation for anomaly detection. In *European Conference on Computer Vision*, pages 438–454. Springer, 2024. 3
- [19] Sungwook Lee, Seunghyun Lee, and Byung Cheol Song. Cfa: Coupled-hypersphere-based feature adaptation for target-oriented anomaly localization. *IEEE Access*, 10: 78446–78454, 2022. 3, 7
- [20] Chun-Liang Li, Kihyuk Sohn, Jinsung Yoon, and Tomas Pfister. Cutpaste: Self-supervised learning for anomaly detection and localization. In *CVPR*, pages 9664–9674, 2021. 2, 3
- [21] Pengzhi Li, Qiang Nie, Ying Chen, Xi Jiang, Kai Wu, Yuhuan Lin, Yong Liu, Jinlong Peng, Chengjie Wang, and Feng Zheng. Tuning-free image customization with image and text guidance. In *Proceedings of the European Conference on Computer Vision*, 2024. 2
- [22] Xurui Li, Ziming Huang, Feng Xue, and Yu Zhou. Musc: Zero-shot industrial anomaly classification and segmentation with mutual scoring of the unlabeled images. *arXiv preprint arXiv:2401.16753*, 2024. 1, 3, 7
- [23] 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 *ICME*, pages 1–6. IEEE, 2021. 2, 3, 6, 1
- [24] Zhikang Liu, Yiming Zhou, Yuansheng Xu, and Zilei Wang. Simplenet: A simple network for image anomaly detection and localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20402–20411, 2023. 3
- [25] Sangwoo Mo, Minsu Cho, and Jinwoo Shin. Freeze the discriminator: a simple baseline for fine-tuning gans. *arXiv preprint arXiv:2002.10964*, 2020. 2
- [26] 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. 2, 3, 6, 1
- [27] Utkarsh Ojha, Yijun Li, Jingwan Lu, Alexei A Efros, Yong Jae Lee, Eli Shechtman, and Richard Zhang. Few-shot image generation via cross-domain correspondence. In *CVPR*, pages 10743–10752, 2021. 6, 1, 2

- [28] Guansong Pang, Choubo Ding, Chunhua Shen, and Anton van den Hengel. Explainable deep few-shot anomaly detection with deviation networks. *arXiv preprint arXiv:2108.00462*, 2021. 7
- [29] Jinlong Peng, Zekun Luo, Liang Liu, and Boshen Zhang. Frih: Fine-grained region-aware image harmonization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4478–4486, 2024. 2
- [30] Xuebin Qin, Zichen Zhang, Chenyang Huang, Masood Dehghan, Osmar R Zaiane, and Martin Jagersand. U2-net: Going deeper with nested u-structure for salient object detection. *Pattern recognition*, 106:107404, 2020. 5, 1
- [31] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 4
- [32] 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 *CVPR*, pages 13576–13586, 2022. 7
- [33] 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*, pages 10684–10695, 2022. 2, 4
- [34] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention—MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015. 4
- [35] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. 2
- [36] Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter Gehler. Towards total recall in industrial anomaly detection. In *CVPR*, pages 14318–14328, 2022. 1, 3, 7
- [37] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *CVPR*, pages 22500–22510, 2023. 2, 4
- [38] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016. 1
- [39] Thomas Schlegl, Philipp Seeböck, Sebastian M Waldstein, Georg Langs, and Ursula Schmidt-Erfurth. f-anogan: Fast unsupervised anomaly detection with generative adversarial networks. *Medical image analysis*, 54:30–44, 2019. 3
- [40] Chengjie Wang, Chengming Xu, Zhenye Gan, Yuxi Li, Jianlong Hu, Wenbing Zhu, and Lizhuang Ma. Pspu: Enhanced positive and unlabeled learning by leveraging pseudo supervision. In *2024 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2024. 3
- [41] Chengjie Wang, Wenbing Zhu, Bin-Bin Gao, Zhenye Gan, Jiangning Zhang, Zhihao Gu, Shuguang Qian, Mingang Chen, and Lizhuang Ma. Real-iad: A real-world multi-view dataset for benchmarking versatile industrial anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22883–22892, 2024.
- [42] Chengjie Wang, Xi Jiang, Bin-Bin Gao, Zhenye Gan, Yong Liu, Feng Zheng, and Lizhuang Ma. Softpatch+: Fully unsupervised anomaly classification and segmentation. *Pattern Recognition*, 161:111295, 2025. 3
- [43] Jinglong Wang, Xiawei Li, Jing Zhang, Qingyuan Xu, Qin Zhou, Qian Yu, Lu Sheng, and Dong Xu. Diffusion model is secretly a training-free open vocabulary semantic segmenter. *arXiv preprint arXiv:2309.02773*, 2023. 5
- [44] Yue Wang, Jinlong Peng, Jiangning Zhang, Ran Yi, Yabiao Wang, and Chengjie Wang. Multimodal industrial anomaly detection via hybrid fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8032–8041, 2023. 3
- [45] Ran Yi, Teng Hu, Mengfei Xia, Yizhe Tang, and Yong-Jin Liu. Feditnet++: Few-shot editing of latent semantics in gan spaces with correlated attribute disentanglement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 2
- [46] Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj. Draem—a discriminatively trained reconstruction embedding for surface anomaly detection. In *ICCV*, pages 8330–8339, 2021. 2, 3, 6, 7
- [47] 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*, pages 2524–2534, 2021. 2, 3, 6, 1
- [48] Hui Zhang, Zuxuan Wu, Zheng Wang, Zheneng Chen, and Yu-Gang Jiang. Prototypical residual networks for anomaly detection and localization. In *CVPR*, pages 16281–16291, 2023. 1, 3, 6, 7
- [49] Lvmin Zhang and Maneesh Agrawala. Transparent image layer diffusion using latent transparency. *arXiv preprint arXiv:2402.17113*, 2024. 4
- [50] Shengyu Zhao, Zhijian Liu, Ji Lin, Jun-Yan Zhu, and Song Han. Differentiable augmentation for data-efficient gan training. *NIPS*, 33:7559–7570, 2020. 1
- [51] Jingyuan Zhu, Huimin Ma, Jiansheng Chen, and Jian Yuan. High-quality and diverse few-shot image generation via masked discrimination. *IEEE Transactions on Image Processing*, 2024. 2

# Dual-Interrelated Diffusion Model for Few-Shot Anomaly Image Generation

## Supplementary Material

### 6. Implementation Details

#### 6.1. Training Details

We train a set of model parameters for each anomaly type. The model requires 5,000 epochs for training, which takes approximately 4.5 hours on an NVIDIA V100 32GB GPU. With BCM, the training step requires only 2,000 epochs. The batch size is set to 4, the learning rate is 0.000005, and the rank of LoRA is 32.

During training, we utilize random flipping for data augmentation. For the “Background Compensation Module”, we have applied it to all categories that involve backgrounds. Among them, categories bottle, pill and toothbrush have witnessed a highly significant improvement. The improvement in the other several categories is rather limited since they can already be generated with a high level of quality.

#### 6.2. Inference Details

During inference without BCM, we only need to input a set of prompts: “a vfx with sks” and “sks”. This process generates a set of anomaly images along with the corresponding anomaly part images. We generate 1000 image pairs with a resolution of 512×512 for each anomaly. Specifically, the num\_inference\_steps is set to 50, and the guidance\_scale is set to 2.5. Notably, it takes 15 seconds to generate each pair of images.

#### 6.3. Mask Generation

We employ U<sup>2</sup>-Net [30] to segment the anomaly part image and obtain the corresponding mask. Based on our observations, this mask is entirely accurate.

### 7. More Ablation Studies

We present comprehensive pixel-level and image-level results for downstream anomaly detection in Tables 6 and 7. The term “dual-interrelated diffusion” refers to the utilization of the dual-interrelated model framework, where the type name such as “cable” is employed as a prompt. The notation “+ prompt” indicates the replacement of the type name with “vfx” and “sks”. Additionally, “+BCM” signifies the incorporation of the Background Compensation Module, which is specifically applied to the categories of bottle, grid, hazelnut, pill, and screw. It can be observed that the prompt we designed outperforms the use of category names, with the exception of the toothbrush category. However, the gap between the toothbrush category and the prompt can be effectively bridged by the BCM module.

### 8. More Qualitative Experiments

We conducted a comprehensive comparison between our generated results and those of existing anomaly image generation methods, with the results presented in Fig. 7. It is evident that the diversity of anomalies generated by Crop&Paste [23] is limited. The results from DiffAug [50] exhibit overfitting. The generated outcomes from CDC [27] lack realism, often resulting in distortion, deformation, and other artifacts. SDGAN [26] and Defect-GAN [47] fail to generate masks corresponding to the anomalies, and the authenticity of the generated images is also limited. The masks produced by DFMGAN [6] are not sufficiently aligned, often resulting in the generation of spots or noise. The currently best-performing method, Anomaly-Diffusion [14], solely focuses on learning the anomaly part. Consequently, the generated anomaly data fails to integrate smoothly with the original image. And, this sometimes leads to the situation where anomalies manifest against the backdrop of the image. In contrast, our method not only generates highly realistic and diverse anomaly data but also produces highly aligned corresponding masks.

Among all these methods, DFMGAN and AnomalyDiffusion are currently the two best performers, so we conducted a more detailed visualization comparison of our results with these two methods. Additional visualizations are presented in Fig. 8-14. The left side shows two examples from the training data, while the right side displays the generated image pairs.

### 9. Quantitative Experiments Setting

#### 9.1. Generated Data

In all comparison methods, 1000 sets of data are generated for each subclass for downstream detection tasks.

#### 9.2. Metrics

This section provides supplementary information on the rationale for using these indicators and their definitions.

**For Generation.** General image generation tasks typically use **Fréchet Inception Distance (FID)** [11] to evaluate the difference between the generated data and the real data distribution. However, FID is not reliable in cases of limited anomalous data, as it tends to produce higher scores for overfitted models. Therefore, we utilize the **Inception Score (IS)** [38] as our evaluation metric. The IS does not require training data and quantifies the quality and diversity of the generated images by calculating the negative exponent of the Kullback-Leibler (KL) divergence between the

Category	dual-interrelated diffusion			+prompt			+prompt +BCM		
	AUC-P	AP-P	$F_1$ -P	AUC-P	AP-P	$F_1$ -P	AUC-P	AP-P	$F_1$ -P
bottle	96.4	74.2	69.7	98.4	88.8	77.1	<b>99.5</b>	<b>93.4</b>	<b>85.7</b>
cable	95.7	74.1	68.8	<b>97.5</b>	<b>82.6</b>	<b>76.9</b>	<b>97.5</b>	<b>82.6</b>	<b>76.9</b>
capsule	97.8	54.8	54.3	<b>99.5</b>	<b>73.2</b>	<b>67.0</b>	<b>99.5</b>	<b>73.2</b>	<b>67.0</b>
carpet	99.4	86.7	77.9	<b>99.4</b>	<b>89.1</b>	<b>80.2</b>	<b>99.4</b>	<b>89.1</b>	<b>80.2</b>
grid	95.8	36.2	39.8	98.5	57.2	54.9	<b>98.5</b>	<b>57.2</b>	<b>54.9</b>
hazelnut	99.5	94.8	89.9	<b>99.8</b>	96.5	91.5	<b>99.8</b>	<b>97.7</b>	<b>92.8</b>
leather	98.4	79.1	70.1	<b>99.9</b>	<b>88.8</b>	<b>78.8</b>	<b>99.9</b>	<b>88.8</b>	<b>78.8</b>
metal_nut	98.8	94.4	89.1	<b>99.6</b>	<b>98.0</b>	<b>93.0</b>	<b>99.6</b>	<b>98.0</b>	<b>93.0</b>
pill	89.6	38.1	31.2	98.4	86.9	78.2	<b>99.6</b>	<b>95.8</b>	<b>89.2</b>
screw	97.7	48.9	47.9	97.7	55.15	72.8	<b>98.1</b>	<b>57.1</b>	<b>56.1</b>
tile	99.1	91.0	80.8	<b>99.7</b>	<b>97.1</b>	<b>91.0</b>	<b>99.7</b>	<b>97.1</b>	<b>91.0</b>
toothbrush	98.2	65.2	67.1	97.2	62.7	64.0	<b>98.2</b>	<b>68.3</b>	<b>68.6</b>
transistor	94.9	78.2	73.1	<b>98.0</b>	<b>86.7</b>	<b>79.6</b>	<b>98.0</b>	<b>86.7</b>	<b>79.6</b>
wood	98.6	87.3	75.9	<b>99.4</b>	<b>91.6</b>	<b>83.8</b>	<b>99.4</b>	<b>91.6</b>	<b>83.8</b>
zipper	98.4	82.2	72.5	<b>99.6</b>	<b>90.7</b>	<b>82.7</b>	<b>99.6</b>	<b>90.7</b>	<b>82.7</b>
Average	97.22	72.35	67.21	98.8	83.0	78.1	<b>99.1</b>	<b>84.5</b>	<b>78.8</b>

Table 6. Ablaiton Study: comparison on pixel-level anomaly localization on the MVTec dataset by training a U-Net on our model’s generated data using different settings.

Category	dual-interrelated diffusion			+prompt			+prompt +BCM		
	AUC-P	AP-P	$F_1$ -P	AUC-I	AP-I	$F_1$ -I	AUC-P	AP-P	$F_1$ -P
bottle	98.0	99.2	96.4	98.7	98.0	98.9	<b>100</b>	<b>100</b>	<b>100</b>
cable	92.3	94.5	85.1	<b>97.7</b>	<b>98.3</b>	<b>94.2</b>	<b>97.7</b>	<b>98.3</b>	<b>94.2</b>
capsule	81.9	93.5	88.9	<b>97.6</b>	<b>99.2</b>	<b>95.8</b>	<b>97.6</b>	<b>99.2</b>	<b>95.8</b>
carpet	96.7	98.8	95.7	<b>99.8</b>	<b>99.9</b>	<b>99.1</b>	<b>99.8</b>	<b>99.9</b>	<b>99.1</b>
grid	97.2	98.6	95.0	<b>99.5</b>	<b>99.7</b>	<b>97.6</b>	<b>99.5</b>	<b>99.7</b>	<b>97.6</b>
hazelnut	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>
leather	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>
metal_nut	97.7	99.3	97.6	<b>99.7</b>	<b>99.9</b>	<b>99.2</b>	<b>99.7</b>	<b>99.9</b>	<b>99.2</b>
pill	87.1	96.3	91.2	92.0	97.8	93.6	<b>95.8</b>	<b>99.0</b>	<b>95.8</b>
screw	83.5	90.1	84.1	86.6	94.2	86.1	<b>87.8</b>	<b>95.0</b>	<b>87.2</b>
tile	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>
toothbrush	97.9	98.8	94.7	97.6	98.5	93.9	<b>99.5</b>	<b>99.7</b>	<b>97.5</b>
transistor	92.8	92.3	89.4	<b>95.1</b>	<b>93.7</b>	<b>90.1</b>	<b>95.1</b>	<b>93.7</b>	<b>90.1</b>
wood	99.3	99.7	97.6	<b>100</b>	<b>99.9</b>	<b>100</b>	<b>100</b>	<b>99.9</b>	<b>100</b>
zipper	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>
Average	94.9	97.4	94.38	97.6	98.6	96.5	<b>99.8</b>	<b>98.9</b>	<b>99.8</b>

Table 7. Ablaiton Study: comparison on image-level anomaly localization on the MVTec dataset by training a U-Net on our model’s generated data using different settings.

edge distribution of the generated images and the conditional distribution of the class labels predicted by the Inception model. A higher IS score indicates better quality and diversity in the generated images.

In addition, we use **Intra-cluster Pairwise LPIPS Distance (IC-LPIPS)** [27] to measure the diversity of the generated data. This method clusters the images into k groups based on the LPIPS distance to k target samples and then computes the average mean LPIPS distances to the corresponding target samples within each cluster. Higher IC-LPIPS scores indicate better diversity.

**For Anomaly Inspection.** We use the **Area Under the**

**Receiver Operating Characteristic (AUROC), Average Precision (AP), and  $F_1$ -max** to measure the performance of the inspection following the general anomaly inspection task.

### 9.3. Anomaly Inspection Detail

In the downstream task of anomaly detection, we employ a simple U-Net [35] architecture. To mitigate the effects of randomness, we train the model three times and select the best result as the final outcome.

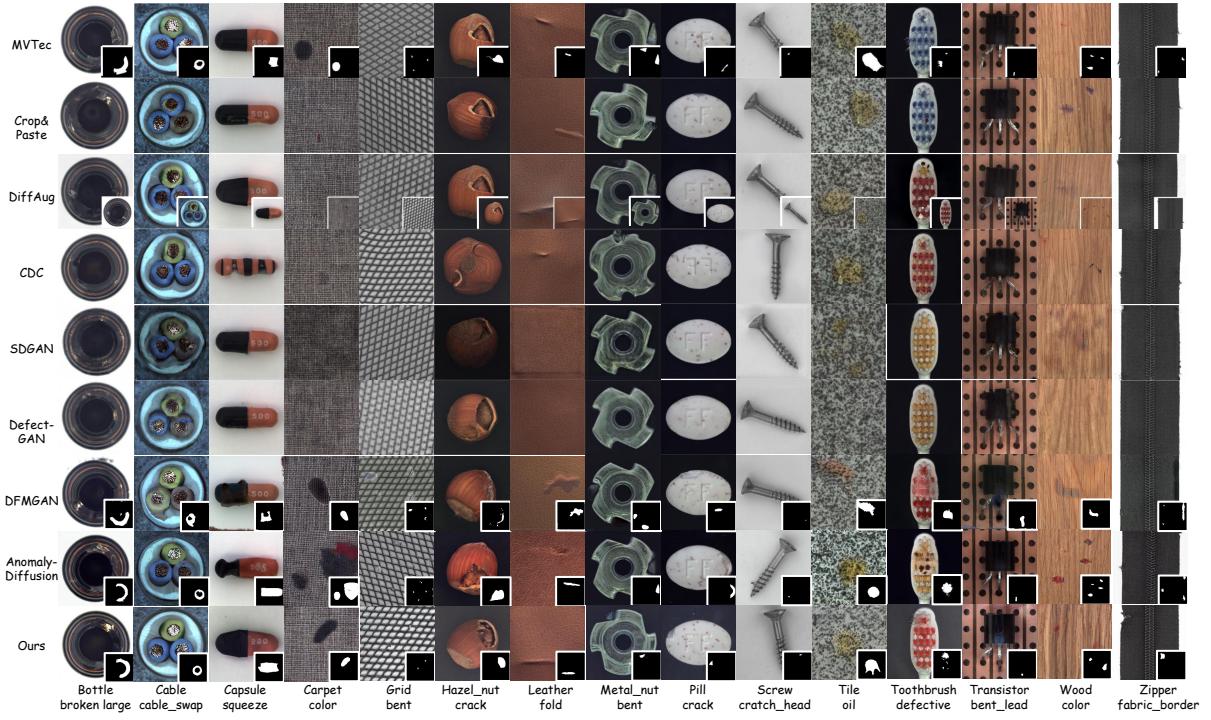


Figure 7. Comparison on the generation results on MVTec.

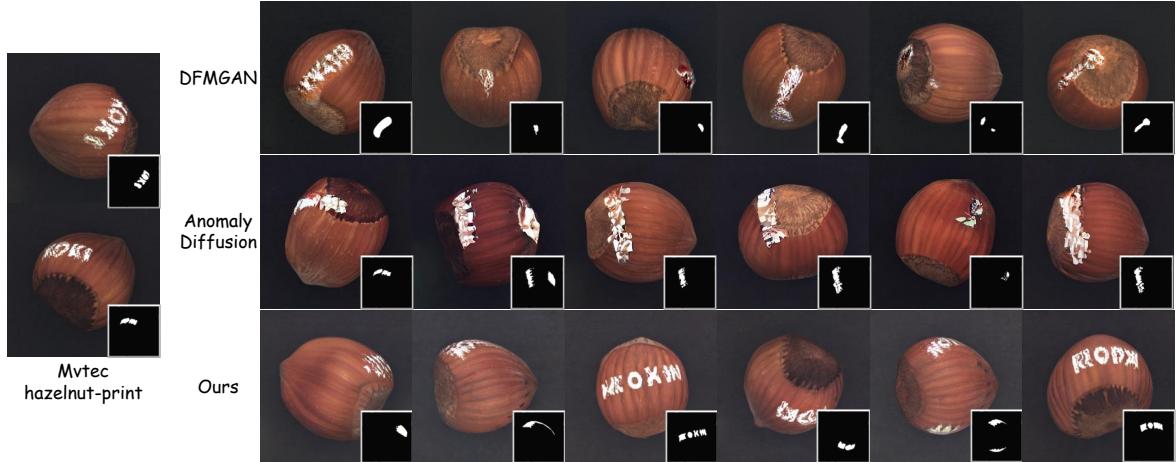


Figure 8. Comparison on the type of hazelnut-print. DFMGAN and AnomalyDiffusion struggle to generate realistic anomalies, particularly in the print class, where the anomalous regions consist of strings of letters. In contrast, our method successfully generates both the shape of the letters and the corresponding mask that aligns with their contours.

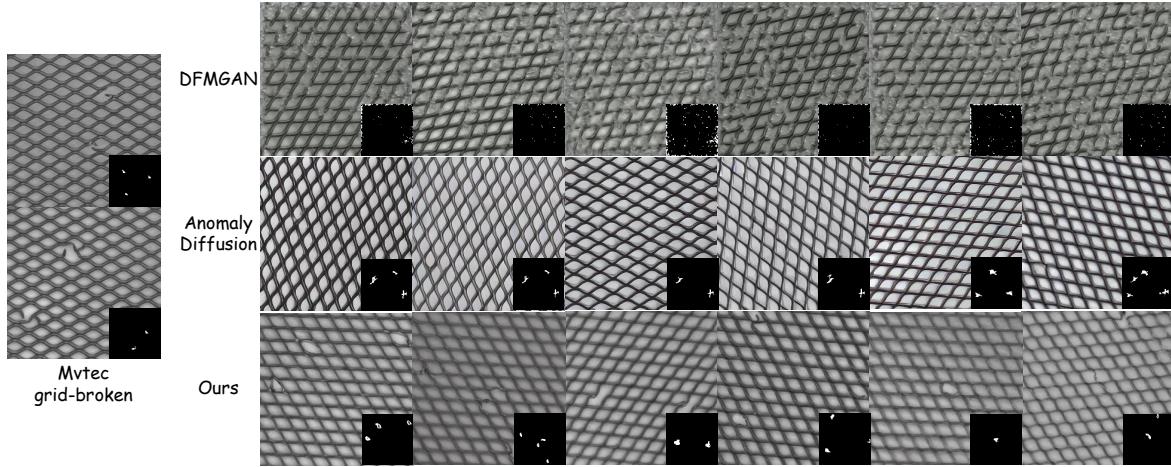


Figure 9. Comparison on the type of grid-broken. For this specific type of small, structure-related anomaly, the images generated by DFMGAN are of poor quality, and AnomalyDiffusion fails to produce any anomalies. In contrast, our method generates highly realistic and effective anomaly images.

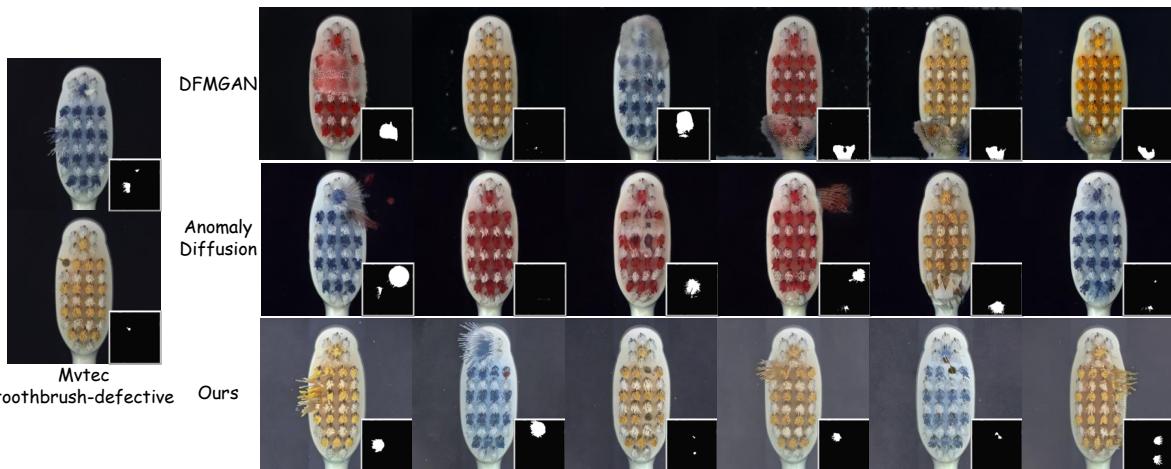


Figure 10. Comparison on the type of toothbrush-defective. The anomalies generated by DFMGAN lack realism, while those produced by AnomalyDiffusion are detached from the main object. Additionally, the generated anomalies, such as holes and bristles of toothbrushes, are mixed. In contrast, although there are some differences in background color, the generated anomalies by our model are fully consistent with real-world scenarios. Furthermore, the background issues do not impact the effectiveness of anomaly detection in downstream tasks.

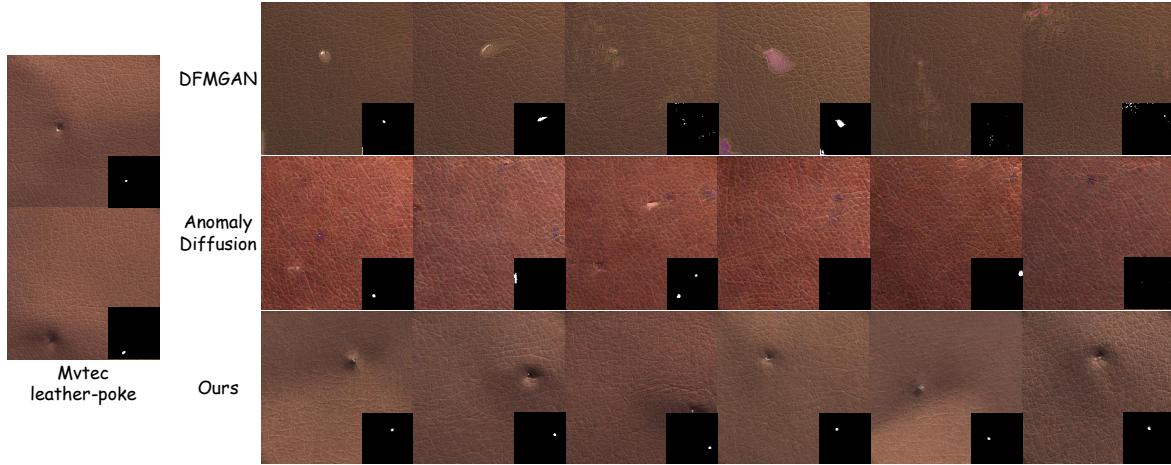


Figure 11. Comparison on the type of leather-poke. The anomalies generated by AnomalyDiffusion are slightly better than those produced by DFMGAN, however, there is a noticeable color difference in the leather. In contrast, our method achieves good results in both aspects.



Figure 12. Comparison on the type of capsule-scratch. For scratches, a relatively minor type of anomaly, neither DFMGAN nor Anomaly-Diffusion can generate effective results. In contrast, our method not only produces realistic anomalies but also demonstrates a good variety.

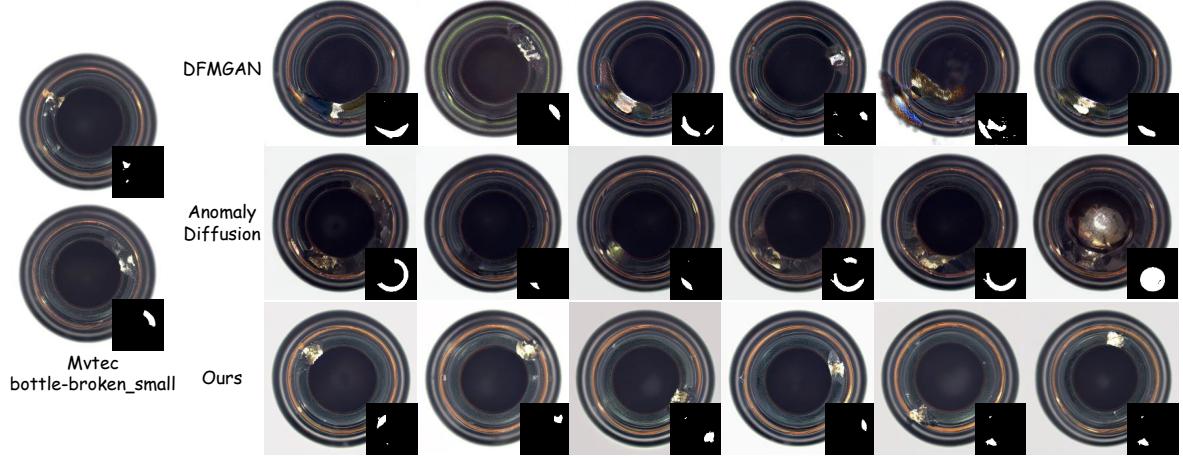


Figure 13. Comparison on the type of bottle-broken\_small. This type of anomaly refers to a small blemish around the edge of a bottle, while broken\_large indicates a larger blemish in the same area. The quality of the image generated by DFMGAN is limited, and the mask are not properly aligned. while the abnormal position generated by AnomalyDiffusion sometimes is not correct, and the shape does not belong to the type of broken\_small, but more like broken\_large. Our method, however, achieves good results in both position and shape.

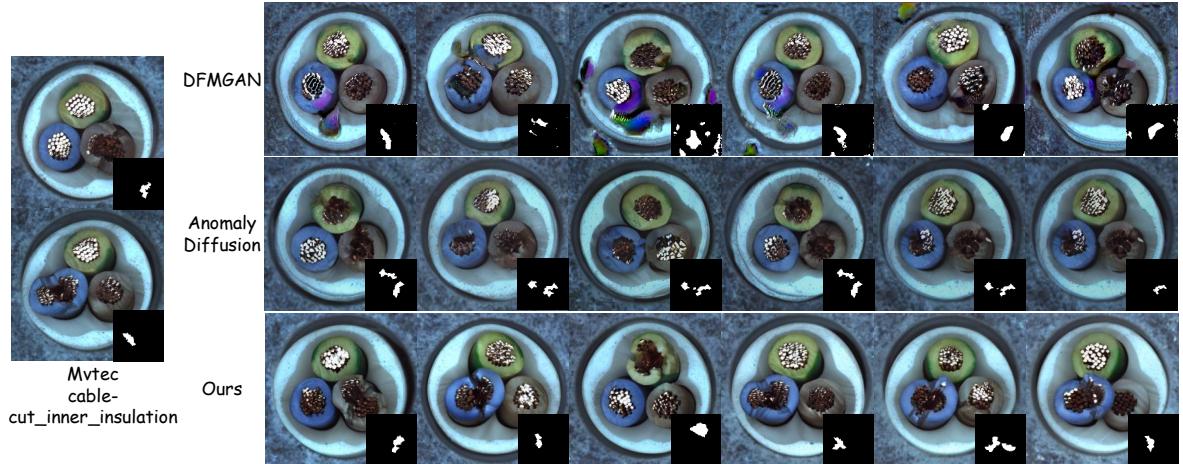


Figure 14. Comparison on the type of cable-cut\_inner\_insulation. It is evident that neither DFMGAN nor AnomalyDiffusion can generate realistic anomalies, and the overall quality of the images produced by DFMGAN is subpar. In contrast, our method successfully generates realistic and diverse abnormal data.