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

Supplementary Material

1. Implementation Details

1.1. 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.

1.2. 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.

1.3. Mask Generation

We employ U^2-Net [7] to segment the anomaly part image and obtain the corresponding mask. Based on our observations, this mask is entirely accurate.

2. More Ablation Studies

We present comprehensive pixel-level and image-level results for downstream anomaly detection in Tables 1 and 2. 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.

3. 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. 1. It is evident that the diversity of anomalies generated by Crop&Paste [4] is limited. The results from DiffAug [11] exhibit overfitting. The generated outcomes from CDC [6] lack realism, often resulting in distortion, deformation, and other artifacts. SDGAN [5] and Defect-GAN [10] fail to generate masks corresponding to the anomalies, and the authenticity of the generated images is also limited. The masks produced by DFMGAN [1] are not sufficiently aligned, often resulting in the generation of spots or noise. The currently best-performing method, AnomalyDiffusion [3], 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. 2-8. The left side shows two examples from the training data, while the right side displays the generated image pairs.

4. Quantitative Experiments Setting

4.1. Generated Data

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

4.2. 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)** [2] 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)** [9] 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	**99.5**	**93.4**	**85.7**
cable	95.7	74.1	68.8	**97.5**	**82.6**	**76.9**	**97.5**	**82.6**	**76.9**
capsule	97.8	54.8	54.3	**99.5**	**73.2**	**67.0**	**99.5**	**73.2**	**67.0**
carpet	99.4	86.7	77.9	**99.4**	**89.1**	**80.2**	**99.4**	**89.1**	**80.2**
grid	95.8	36.2	39.8	98.5	57.2	54.9	**98.5**	**57.2**	**54.9**
hazelnut	99.5	94.8	89.9	**99.8**	96.5	91.5	**99.8**	**97.7**	**92.8**
leather	98.4	79.1	70.1	**99.9**	**88.8**	**78.8**	**99.9**	**88.8**	**78.8**
metal_nut	98.8	94.4	89.1	**99.6**	**98.0**	**93.0**	**99.6**	**98.0**	**93.0**
pill	89.6	38.1	31.2	98.4	86.9	78.2	**99.6**	**95.8**	**89.2**
screw	97.7	48.9	47.9	97.7	55.15	72.8	**98.1**	**57.1**	**56.1**
tile	99.1	91.0	80.8	**99.7**	**97.1**	**91.0**	**99.7**	**97.1**	**91.0**
toothbrush	98.2	65.2	67.1	97.2	62.7	64.0	**98.2**	**68.3**	**68.6**
transistor	94.9	78.2	73.1	**98.0**	**86.7**	**79.6**	**98.0**	**86.7**	**79.6**
wood	98.6	87.3	75.9	**99.4**	**91.6**	**83.8**	**99.4**	**91.6**	**83.8**
zipper	98.4	82.2	72.5	**99.6**	**90.7**	**82.7**	**99.6**	**90.7**	**82.7**
Average	97.22	72.35	67.21	98.8	83.0	78.1	**99.1**	**84.5**	**78.8**

Table 1. 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	**100**	**100**	**100**
cable	92.3	94.5	85.1	**97.7**	**98.3**	**94.2**	**97.7**	**98.3**	**94.2**
capsule	81.9	93.5	88.9	**97.6**	**99.2**	**95.8**	**97.6**	**99.2**	**95.8**
carpet	96.7	98.8	95.7	**99.8**	**99.9**	**99.1**	**99.8**	**99.9**	**99.1**
grid	97.2	98.6	95.0	**99.5**	**99.7**	**97.6**	**99.5**	**99.7**	**97.6**
hazelnut	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**
leather	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**
metal_nut	97.7	99.3	97.6	**99.7**	**99.9**	**99.2**	**99.7**	**99.9**	**99.2**
pill	87.1	96.3	91.2	92.0	97.8	93.6	**95.8**	**99.0**	**95.8**
screw	83.5	90.1	84.1	86.6	94.2	86.1	**87.8**	**95.0**	**87.2**
tile	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**
toothbrush	97.9	98.8	94.7	97.6	98.5	93.9	**99.5**	**99.7**	**97.5**
transistor	92.8	92.3	89.4	**95.1**	**93.7**	**90.1**	**95.1**	**93.7**	**90.1**
wood	99.3	99.7	97.6	**100**	**99.9**	**100**	**100**	**99.9**	**100**
zipper	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**
Average	94.9	97.4	94.38	97.6	98.6	96.5	**99.8**	**98.9**	**99.8**

Table 2. 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)** [6] 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.

4.3. Anomaly Inspection Detail

In the downstream task of anomaly detection, we employ a simple U-Net [8] architecture. To mitigate the effects of randomness, we train the model three times and select the best result as the final outcome.

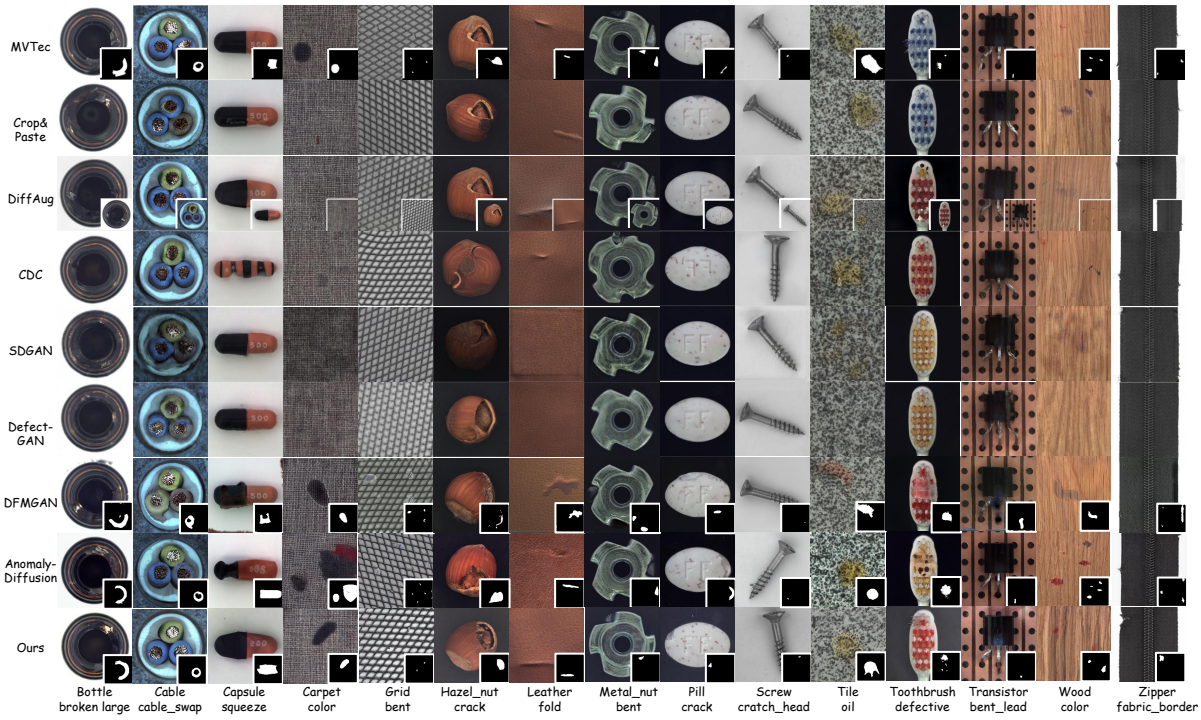

Figure 1. Comparison on the generation results on MVTec.

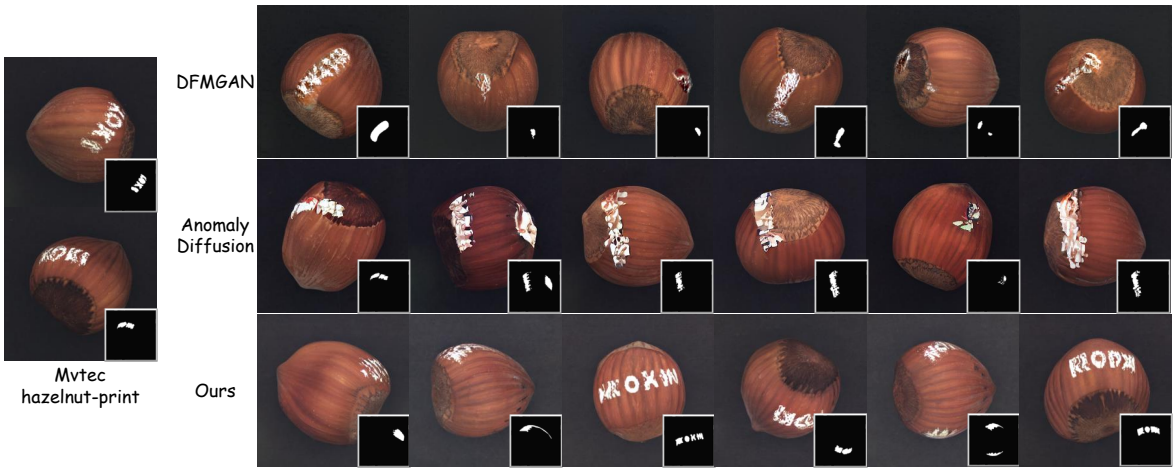

Figure 2. Comparison on the type of hazelnut-print. DFMGAN and AnomalyDiffusion struggle to generate realistic anomalies, particularly in the print class, where the anomalous regions consist of strings of letters. In contrast, our method successfully generates both the shape of the letters and the corresponding mask that aligns with their contours.

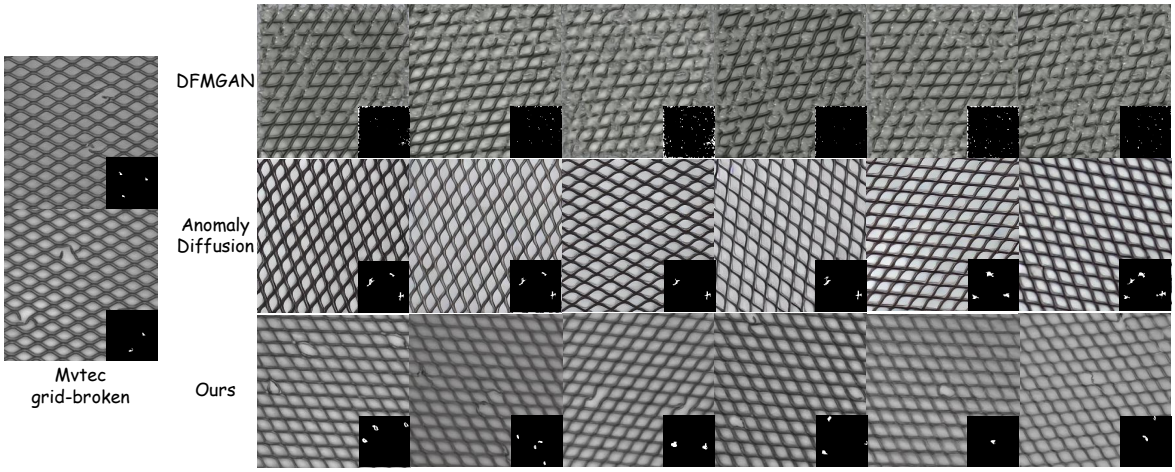

Figure 3. Comparison on the type of grid-broken. For this specific type of small, structure-related anomaly, the images generated by DFMGAN are of poor quality, and AnomalyDiffusion fails to produce any anomalies. In contrast, our method generates highly realistic and effective anomaly images.

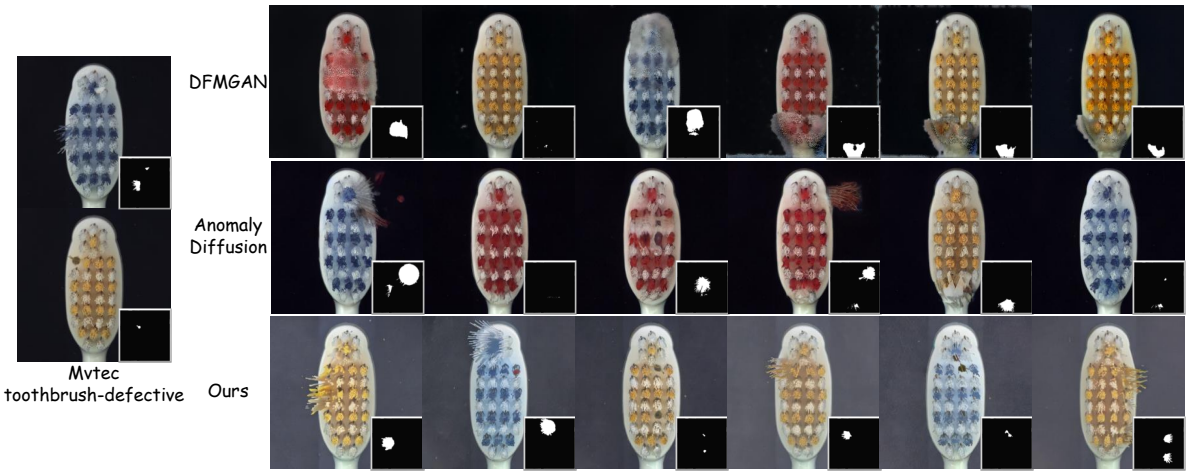

Figure 4. Comparison on the type of toothbrush-defective. The anomalies generated by DFMGAN lack realism, while those produced by AnomalyDiffusion are detached from the main object, Additionally, the generated anomalies, such as holes and bristles of toothbrushes, are mixed. In contrast, although there 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.

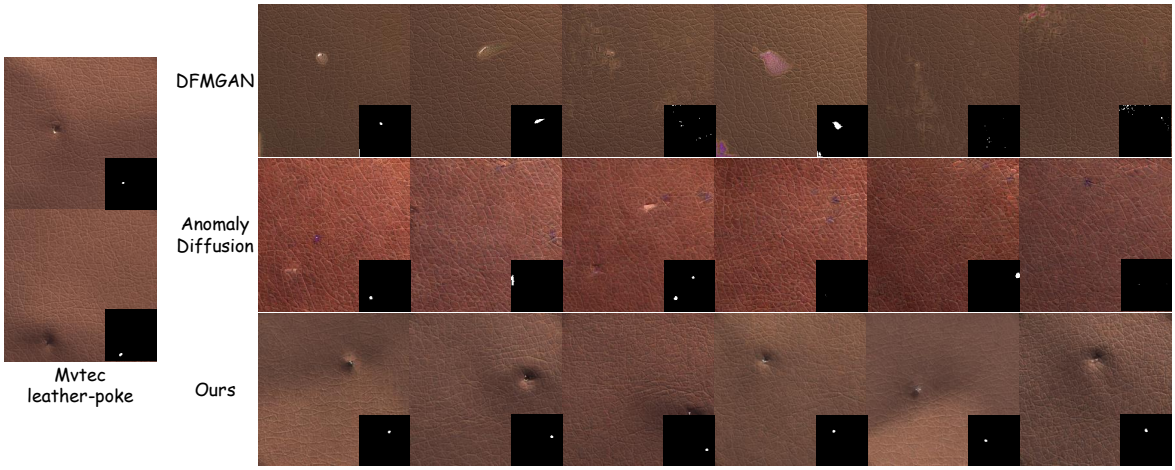

Figure 5. Comparison on the type of leather-poke. The anomalies generated by AnomalyDiffusion are slightly better than those produced by DFMGAN, however, there is a noticeable color difference in the leather. In contrast, our method achieves good results in both aspects.

Figure 6. Comparison on the type of capsule-scratch. For scratches, a relatively minor type of anomaly, neither DFMGAN nor AnomalyDiffusion can generate effective results. In contrast, our method not only produces realistic anomalies but also demonstrates a good variety.

Figure 7. 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.

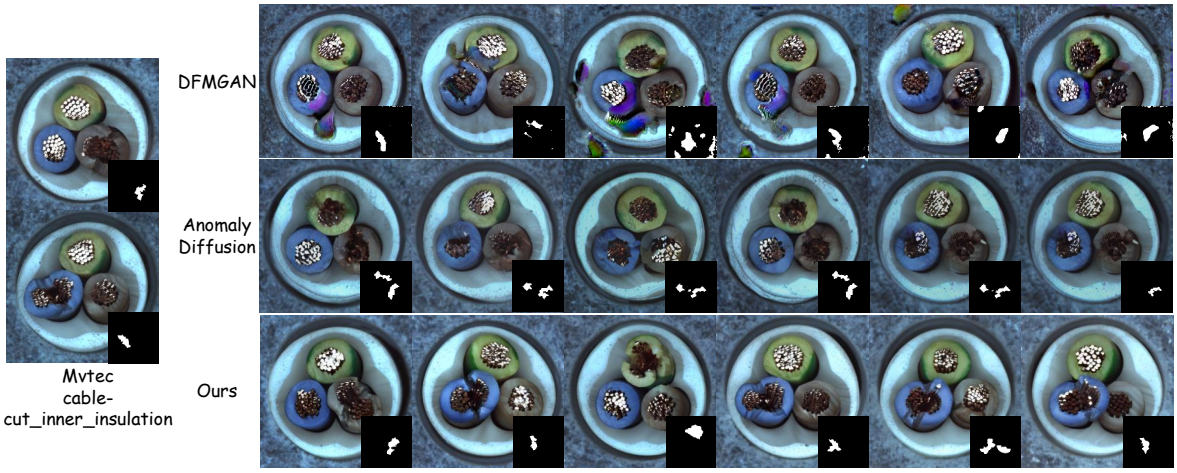

Figure 8. Comparison on the type of cable-cut_inner_insulation. It is evident that neither DFMGAN nor AnomalyDiffusion can generate realistic anomalies, and the overall quality of the images produced by DFMGAN is subpar. In contrast, our method successfully generates realistic and diverse abnormal data.